On the Benefits of Distributed Populations for Noisy Optimization

Dirk V. Arnold and Hans-Georg Beyer
Department of Computer Science XI
University of Dortmund
44221 Dortmund, Germany
{arnold,beyer}@ls11.cs.uni-dortmund.de

Abstract

While in the absence of noise, no improvement in local performance can be gained from retaining but the best candidate solution found so far, it has been shown experimentally that in the presence of noise, operating with a non-singular population of candidate solutions can have a marked and positive effect on the local performance of evolution strategies. So as to determine the reasons for the improved performance, we study the evolutionary dynamics of the (μ, λ)-ES in the presence of noise. Considering a simple, idealized environment, a moment-based approach that utilizes recent results involving concomitants of selected order statistics is developed. This approach yields an intuitive explanation for the performance advantage of multi-parent strategies in the presence of noise. It is then shown that the idealized dynamic process considered does bear relevance to optimization problems in high-dimensional search spaces.

Keywords: Evolution strategies, (μ, λ)-ES, distributed populations, population variance, Gaussian noise, noise-to-signal ratio.

1 Introduction

Evolutionary algorithms (EAs) are frequently recommended for optimization in the presence of noise. Underlying this recommendation is a good track record of evolutionary optimization strategies on noisy, real-world optimization problems along with the vague idea that using a population of candidate solutions that is distributed in search space should make EAs particularly robust and insensitive to the effects of noise. Empirical research by Nissen and Propach [11] seems to support this idea. However, little is known as to where exactly the benefits of distributed populations in noisy environments stem from. In the realm of genetic algorithms (GAs), some results have been found by Miller and Goldberg [10] and by Rattray and Shapiro [13]. Based on the building block hypothesis, Miller and Goldberg analyzed the effects of noise on different selection mechanisms for GAs. Rattray and Shapiro investigated the effects of noise on GA performance on the *OneMax* function and on a perceptron learning problem with binary weights. They concluded that the effects of noise can be removed altogether by using a sufficiently large population, where the necessary population size increases exponentially with the noise strength. While both of those studies consider discrete problems, the present paper focuses on the performance of evolution strategies (ES) in continuous search spaces. The mathematical analysis of the performance of a multi-parent ES in the presence of noise would greatly contribute to the understanding of the reasons for the observed robustness of EAs. According to Rechenberg [15], it would be a "little breakthrough" if a law describing the local performance of a multi-parent strategy in a quadratic fitness environment could be found.

In the absence of noise, Beyer [6] presented a moment-based analysis of the performance of the (μ, λ)-ES for spherically symmetric fitness functions. The mathematical difficulties of the analysis were considerable as for $\mu > 1$, the population of candidate solutions that emerges in the course of the evolution is distributed in search space and needs to be modeled. The approach relied on the possibility to characterize the population distribution by a small number of its lower order moments, such as expected value, variance, skewness, and kurtosis. Expansions of the population distribution in terms of derivatives of the normal distribution were employed. Approximations to the lower-order central moments of the distribution and subsequently to the progress rate were obtained by imposing "self-consistency conditions" and solving a resulting system of equations. The results that were obtained are quite accurate even if only moments up to the third order are considered. A main result of Beyer's analysis, which was also stated by Rechenberg [15], is the observation that on the noise-free sphere the performance of the (μ, λ)-ES with $\mu > 1$ is never superior to that of the $(1, \lambda)$-ES, and that thus no benefits can be gained from retaining any but the best candidate solution generated. However, Rechenberg [15] also provided empirical evidence that this is not true in the presence of noise. Simple computer experiments can be used to demonstrate that for the very same fitness function, significant speed-up factors over the $(1, \lambda)$-ES can be achieved by retaining more than just the (seemingly) best candidate solution if there is noise present.

First steps towards the analysis of the behavior of the (μ, λ)-ES in the presence of noise were taken by Arnold and Beyer [2]. In that reference, a quality gain law for the (μ, λ)-ES on a noisy linear fitness function was derived and its implications were studied. However, the variance of the population of candidate solutions appeared as a factor in that quality gain law, and an attempt to obtain an analytical expression for that variance using a normal approximation to the distribution of candidate solutions yielded unacceptably inaccurate results. Instead, it was necessary to resort to using values for the population variance that were measured empirically. It was noted that in the absence of noise, considering additional terms in the expansion of the population distribution had led to the greatly improved results reported by Beyer [6], and it was suggested that including those terms might yield a much improved approximation in the presence of noise as well. However, it was also noted that the mathematical difficulties involved in such an approach could be expected to be considerable.

A first attempt to overcome those mathematical difficulties was presented by Arnold [1]. In that reference, noisy order statistics were introduced and expected values of samples of selected offspring candidate solutions were computed. Moments of the third order as well as some fourth order terms were included in the analysis. By formulating "self-consistency conditions" analogous to those of Beyer [6], fairly accurate estimates of the quality gain and of the population variance of the (μ, λ)-ES in the presence of noise were obtained. The present paper both extends on those results by improving their accuracy by considering all fourth order terms and it increases the accessibility of the argument by using recent results with respect to expected values of sample moments of concomitants of selected order statistics. Moreover, the process of finding population moments is automated entirely by the *Mathematica* program in the appendix.

The organization of the remainder of this paper is as follows. In Section 2, we study the evolutionary dynamics of the (μ, λ)-ES on a linear, one-dimensional fitness function in the presence of Gaussian noise. Moments up to the fourth order need to be considered so as to obtain a satisfactory approximation to the population distribution. In Section 3, an intuitively appealing explanation for the improved performance of multi-parent strategies in the presence of noise is given based on the insights afforded by the analysis from Section 2. Multi-parent strategies are seen to operate under a reduced noise-to-signal ratio as compared to one-parent strategies, thus enabling selection to be more effective. Finally, in Section 4, we discuss the relevance of the findings that have been made in a highly idealized, one-dimensional fitness environment for practical optimization problems. In particular, implications for the sphere model are discussed, and estimates for optimal population sizes and efficiencies of the (μ, λ)-ES on the noisy sphere are determined for high search space dimensionality. Section 5 concludes with a brief summary of the main results.

2 Analysis

Let us consider the behavior of the (μ, λ)-ES on the simple fitness function

$$\begin{aligned} f &: \mathbb{R} \to \mathbb{R} \\ f(x) &= x. \end{aligned} \tag{1}$$

Without loss of generality, we assume that the task at hand is maximization. Even though very simple, this fitness function can serve to reveal scaling properties of the (μ, λ)-ES and will be seen to have implications for more complex optimization problems. In particular, in Section 4 we will see that it can shed light on design decisions that practitioners face, such as the choice of population size.

The (μ, λ)-ES at time step t maintains a population $\{x_1^{(t)}, \ldots, x_\mu^{(t)}\}$ of μ candidate solutions. So as to obtain the population at time step $t+1$, λ new candidate solutions are generated by each time randomly picking one of the $x_i^{(t)}$ and adding a normally distributed mutation vector with zero mean. Note that for the special case of a one-dimensional fitness function, the mutation vector consists of a single component, and that due to the scale invariance of the fitness function defined in Eq. (1) it can without loss of generality be assumed that the variance of that component is unity. The population of candidate solutions at time step $t+1$ consists of those μ of the λ newly generated candidate solutions that score best in terms of the fitness function. As a noise model, we assume that evaluating the fitness function yields a measured fitness value that differs from the ideal fitness value $f(x)$ by an additive, normally distributed term with mean zero and with variance ϑ^2. This assumption of Gaussian white noise is almost universal in the optimization literature. The decision which candidate solutions to select for inclusion in the next time step's population is made based on measured fitness values. The standard deviation ϑ of the noise term is referred to as the noise level.

As Beyer [6], we consider the central moments

$$m_k = \frac{1}{\mu} \sum_{i=1}^{\mu} (x_i - m_1)^k, \qquad k \geq 2, \tag{2}$$

where $m_1 = \sum_{i=1}^{\mu} x_i/\mu$ denotes the mean, as important characteristics of the population of candidate solutions. For the fitness function defined in Eq. (1), after initialization effects have faded the distribution of the central moments of the population will approach a time-invariant limit distribution. The approach pursued in the remainder of this section consists in determining the influence of mutation and selection on the central moments and inferring properties of their distribution. Note that the change of the mean m_1 of the population is an indicator of the progress that the strategy makes in a single time step. Due to the translation variance of the environment, it is possible without loss of generality to assume that at time t, $m_1 = 0$.

The effects of mutation on the central moments are easily determined. When generating an offspring candidate solution, one of the parents is selected at random and a standard normally distributed random variable is added. The distribution of the offspring is thus the convolution of the distribution of the parents and a standard normal distribution. As such, it has variance $s^2 = m_2 + 1$, coefficient of skewness $\gamma_1 = m_3/\sqrt{m_2+1}^3$, and coefficient of kurtosis $\gamma_2 = (m_4 - 3m_2^2)/(m_2+1)^2$. For a thorough introduction to moments, cumulants, and their interrelationship we refer to Stuart and Ord [17].

The effects of selection are considerably harder to characterize than those of mutation. Mathematically speaking, it is necessary to determine expected values of sample moments of concomitants of selected order statistics. Arnold and Beyer [4] recently presented a moment-based solution to the problem. While their analysis assumes a standardized population, the results can be almost trivially generalized to include the case of a population of arbitrary variance and can be summarized for that case as follows.

Suppose that $(X_1, Y_1), \ldots, (X_\lambda, Y_\lambda)$ are λ independent, identically distributed bivariate observations from a continuous population with probability density function $p(x, y) = p(x|y)q(y)$, where the conditional probability density is $p(x|y) = \phi((x-y)/\vartheta)/\vartheta$. Here and in what follows, $\phi(x)$ denotes the probability density function of the standardized normal distribution. The probability density $q(y)$ is assumed to have variance s^2, coefficient of skewness γ_1, and coefficient of kurtosis γ_2. Intuitively, the X_i are the measured fitness values and the Y_i are the corresponding ideal fitness values. To compute central moments of the selected offspring candidate solutions, let $A = (\alpha_1, \ldots, \alpha_\nu)$ be a vector of ν positive integers α_i, $i = 1, \ldots, \nu$, where $1 \leq \nu \leq \mu$. Furthermore, let

$$S_A = \frac{(\mu-\nu)!}{\mu!} \sum Y_{[i_1:\lambda]}^{\alpha_1} \cdots Y_{[i_\nu:\lambda]}^{\alpha_\nu},$$

where the summation ranges over all indices $i_j = \lambda - \nu + 1, \ldots, \lambda$ such that $i_j \neq i_k$ for any $j \neq k$ and therefore over all candidate solutions that are selected to form the population of the following time step. $Y_{[i:\lambda]}$ denotes the ideal fitness value of the offspring with the ith smallest measured fitness and thus the concomitant of the ith order statistic $X_{i:\lambda}$. A good introduction to concomitants of order statistics can be found in David and Nagaraja [9]. It is easily seen by multiplying out Eq. (2) and rearranging terms that sample moments of the selected offspring candidate solutions can be expressed in terms of the S_A. For example,

$$m_1^{(t+1)} = S_1 \tag{3}$$

$$m_2^{(t+1)} = \frac{\mu-1}{\mu}(S_2 - 2S_{11}) \tag{4}$$

$$m_3^{(t+1)} = \frac{(\mu-1)(\mu-2)}{\mu^2}(S_3 - 3S_{21} + 12S_{111}). \tag{5}$$

Similarly,

$$\begin{aligned} m_4^{(t+1)} - 3m_2^{(t+1)^2} = {} & \frac{(\mu-1)(\mu^2-6\mu+6)}{\mu^3}(S_4 - 4S_{31} + 6S_{22}) \\ & - 12\frac{(\mu-1)(\mu-2)(\mu-3)}{\mu^3}(S_{22} - 2S_{211} + 12S_{1111}). \end{aligned} \tag{6}$$

Thus, expected values of the sample moments at time step $t+1$ can be given provided that expected values of the S_A can be obtained. Expected values of the S_A have been found to be expressible as

$$\mathrm{E}[S_A] = s^{\|A\|} \sum_{i=0}^{\nu} \sum_{k \geq 0} \left[\zeta_{i,0}^{(A)}(k) + \frac{\gamma_1}{6}\zeta_{i,1}^{(A)}(k) + \frac{\gamma_2}{24}\zeta_{i,2}^{(A)}(k) + \frac{\gamma_1^2}{36}\zeta_{i,3}^{(A)}(k) + \ldots \right] h_{\mu,\lambda}^{\nu-i,k}, \tag{7}$$

where $\|A\| = \sum_{i=1}^{\nu} \alpha_i$. The coefficients $\zeta_{i,j}^{(A)}(k)$ have been tabulated by Arnold and Beyer [4] and depend on the noise coefficient

$$a = \frac{1}{\sqrt{1 + (\vartheta/s)^2}}$$

only. The coefficients $h_{\mu,\lambda}^{i,k}$ are defined as

$$h_{\mu,\lambda}^{i,k} = (\lambda - \mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty} \mathrm{He}_k(y)[\phi(y)]^{i+1}[\Phi(y)]^{\lambda-\mu-1}[1 - \Phi(y)]^{\mu-i}\mathrm{d}y, \tag{8}$$

where $\Phi(y)$ denotes the cumulative distribution function of the standardized normal distribution and $\mathrm{He}_k(x)$ denotes the kth Hermite polynomial, and can be determined numerically. For example, for $A = (1)$, the expected value of S_A is

$$\begin{aligned}\mathrm{E}[S_1] = s\Bigg(ah_{\mu,\lambda}^{1,0} &+ \frac{\gamma_1}{6}a^2(3 - 2a^2)h_{\mu,\lambda}^{1,1} \\ &+ \frac{\gamma_2}{24}a^3(4 - 3a^2)h_{\mu,\lambda}^{1,2} - \frac{\gamma_1^2}{36}a^5(6 - 5a^2)(h_{\mu,\lambda}^{1,0} + 2h_{\mu,\lambda}^{1,2}) + \dots\Bigg).\end{aligned} \tag{9}$$

Both Edgeworth and Cornish-Fisher expansions have been used in the derivation of Eq. (7). All terms that are neglected and are represented by dots consist of terms of an order higher than the fourth. It can be expected that additional accuracy could be gained by including such terms in the analysis. However, the resulting expressions become very lengthy, and the accuracy of the results is sufficient for our purposes even if only terms up to the fourth order are considered.

Clearly, the central moments of the population are random variables. For the simple fitness function given in Eq. (1), the probability distribution of those random variables tends to a time-invariant limit distribution. For the simple case of infinite noise level (and thus random selection), some lower order moments of that distribution have been computed by Arnold [1]. Unfortunately, the approach pursued there cannot be used if selection is not random. Instead, we follow Beyer [6] who neglected fluctuations of the central moments in an attempt to learn about their expected values. However, rather than considering central moments, we choose to consider cumulants due to their somewhat better numerical properties. Note that for orders up to the third, central moments and cumulants agree, and that the fourth order cumulant can be obtained from the second and fourth order central moments. We postulate that the expected values of the cumulants of the population at time step $t + 1$ agree with those at time step t, i.e. that

$$\left.\begin{aligned}\mathrm{E}\left[m_2^{(t+1)}\right] &= m_2^{(t)} = m_2 \\ \mathrm{E}\left[m_3^{(t+1)}\right] &= m_3^{(t)} = m_3 \\ \mathrm{E}\left[m_4^{(t+1)} - 3{m_2^{(t+1)}}^2\right] &= m_4^{(t+1)} - 3{m_2^{(t+1)}}^2 = m_4 - 3m_2^2.\end{aligned}\right\} \tag{10}$$

Eqs. (10) together with Eqs. (3) through (6) form a system of three equations in the three unknowns m_2, m_3, and m_4 that can be solved numerically. The *Mathematica* program in the appendix can be used to both obtain expected values of the central moments after selection and for solving the resulting system of equations. While the approach does not yield exact results as both fluctuations and higher order moments have been neglected, it will be seen in Section 3 that the values of the moments that are obtained do not differ much from the expected values that are observed empirically.

3 Discussion

In Figure 1, the estimates that have been obtained by solving the system of Eqs. (10) and the resulting estimate for the expected progress given by Eq. (9) are compared with measurements of runs of evolution strategies. While in that figure we have used $\lambda = 40$, the graphs that are obtained for other values of λ are qualitatively similar. Three levels of approximation are included in the figure. The solid lines are the result of considering moments up to the fourth order, the dashed lines result from considering moments only up to the third order, and the dotted lines correspond to a normal approximation that does not consider either skewness or kurtosis of the distribution. It can be seen that the estimates for the progress and the variance of the population that have been obtained do agree quite closely with the values measured empirically provided that moments including those of the fourth order are considered. For $\mu = 1$, solving Eqs. (10) yields the exact result for the progress as the parental population consists of a single point and the distribution of offspring fitness values is normal. The results agree with those found for the $(1, \lambda)$-ES by Beyer [5]. For $\vartheta = 0$, the estimates agree with those derived by Beyer [6], except that fourth order moments are considered here in addition. Note that while in the absence of noise, considering moments up to the third order is sufficient for obtaining fairly accurate results, in the presence of noise both the expected progress and the population variance are severely overestimated unless moments of the fourth order are included in the analysis.

It can be seen from Figure 1 that in the presence of noise, much higher progress than that of the $(1, \lambda)$-ES can be achieved by choosing $\mu > 1$. As customary, we write $c_{\mu,\lambda}(\vartheta)$ for the expected progress $\mathrm{E}[m_1]$ of the (μ, λ)-ES on the fitness function given by Eq. (1). For nonzero noise, the curves in upper graph have a maximum at intermediate values of μ. For example, while the progress of the $(1, 40)$-ES at noise level $\vartheta = 16.0$ is $c_{1,40}(16.0) \approx 0.13$, that of the $(12, 40)$-ES is $c_{12,40}(16.0) \approx 0.46$. That is, the expected progress can be increased by a factor of 3.5 without additional computational costs simply by retaining more than just the seemingly best offspring candidate solution. Rechenberg [15] empirically demonstrated that at higher noise levels even greater speed-up factors can be observed. The upper graph of Figure 1 suggests that for fixed λ, the optimal size of the parental population μ increases with increasing noise strength — a fact that was already observed by Rechenberg [15] —, and that at the same time the progress becomes less sensitive to the choice of truncation ratio μ/λ. Rechenberg speculated that the fact that the variance of the offspring candidate solutions of a (μ, λ)-ES is increased as compared to that of a $(1, \lambda)$-ES operating with the same mutation strength might contribute to the improved performance in the presence of noise of the former strategy.

So as to see what can be learned with regard to the reasons for the speed-up that can be achieved, let us consider Eq. (9). Closer numerical investigation shows that while skewness and kurtosis are essential for obtaining a satisfactory estimate of the variance m_2 of the population, the influence of the terms in Eq. (9) that γ_1 and γ_2 appear in is rather minor. Omitting all but the first term in the parentheses and using $s^2 = m_2 + 1$ yields the approximation

$$\mathrm{E}[m_1] = c_{\mu,\lambda}(\vartheta) \approx \frac{\sqrt{m_2 + 1}}{\sqrt{1 + \vartheta^2/(m_2 + 1)}} h^{1,0}_{\mu,\lambda}, \tag{11}$$

where the coefficient $h^{1,0}_{\mu,\lambda}$ agrees with the coefficient $e^{1,0}_{\mu,\lambda}$ introduced and studied by Beyer [8]. The dependence of the coefficient on μ and λ is illustrated in Figure 2.

The lower graph of Figure 1 suggests that for fixed λ, the variance of the population increases both with increasing size of the parental population μ and with increasing noise level ϑ. In the absence of noise, the progress according to Eq. (11) simply reads $\mathrm{E}[m_1] \approx \sqrt{m_2 + 1} h^{1,0}_{\mu,\lambda}$. When increasing the

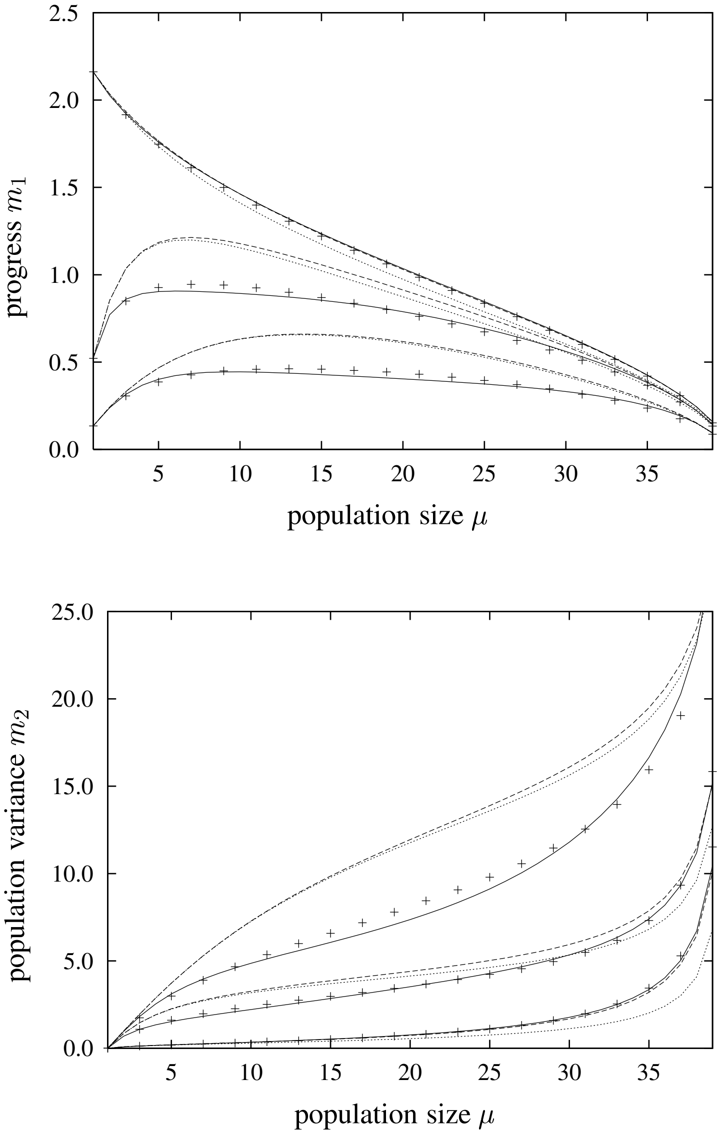

Figure 1: Progress m_1 and population variance m_2 of a (μ, λ)-ES with $\lambda = 40$ as functions of the size of the parental population μ. The lines correspond to, from top to bottom in the upper graph and from bottom to top in the lower graph, noise levels $\vartheta = 0.0$, 4.0, and 16.0. The dotted lines correspond to the normal approximation, the dashed lines take the skewness of the population into account, and the solid lines represent the results obtained when additionally the fourth order moments are considered. The crosses mark data that has been obtained from empirical measurements of runs of evolution strategies.

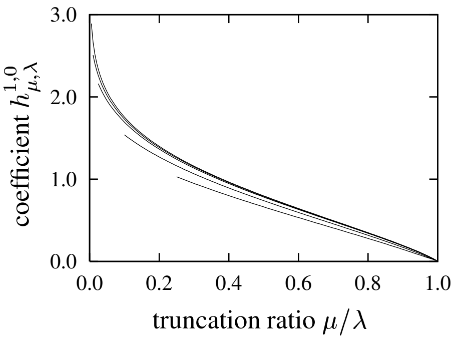

Figure 2: The coefficients $h^{1,0}_{\mu,\lambda}$ defined in Eq. (8) as functions of the truncation ratio μ/λ for different values of λ. The curves correspond to, from bottom to top, $\lambda = 4$, 10, 40, 100, and the limit case $\lambda = \infty$ and are displayed in the range from $1/\lambda$ to 1.0.

size of the parental population μ, the increase in $\sqrt{m_2 + 1}$ due to an increase in the variance m_2 of the population is more than offset by the decrease in $h^{1,0}_{\mu,\lambda}$ if λ remains unchanged. The corresponding curves in the upper graph of Figure 1 decrease monotonically. In the presence of noise, however, the variance of the population also appears under the square root in the denominator where it acts to reduce the weight of the noise-dependent term. As $m_2 + 1$ is the variance of the set of offspring candidate solutions, and as selection is based on the measured fitness of the offspring candidate solutions, the quotient $\vartheta/\sqrt{m_2 + 1}$ is the noise-to-signal ratio of the system. While for the $(1, \lambda)$-ES we have $m_2 = 0$ and the progress decreases with $1/\sqrt{1 + \vartheta^2}$ — a result that is known from Beyer [5] and Rechenberg [15] —, for $\mu > 1$ the noise-to-signal ratio is moderated by the nonzero variance of the population. As we have seen, the effect of this decrease of the noise-to-signal ratio can outweigh the decrease in $\sqrt{m_2 + 1}h^{1,0}_{\mu,\lambda}$ that results from an increase of μ. Eq. (11) thus not only provides a quantitative background for Rechenberg's speculation, but it also demonstrates that at least for the simple fitness function given by Eq. (1) the increased variance of the offspring is the *only* reason for the improved performance of the (μ, λ)-ES as compared with the $(1, \lambda)$-ES. In the next section, we will see that the relevance of the results obtained for that simple fitness function is by no means limited to that function alone, but that the results have important implications for optimization in high-dimensional search spaces.

4 The Sphere

While the objective function introduced in Eq. (1) and studied in Sections 2 and 3 is of a simplicity that makes it appear to be far removed from any practical optimization problem, it does have important implications for all cases in which the fitness function can effectively be linearized locally. While this is frequently possible for one-parent strategies that do not employ a large step length, it is not clear whether linearization does not introduce too large an error if the population is distributed in search space. However, by considering the sphere model — the most frequently studied objective function in the realm of ES — we will see that in high-dimensional search spaces, the variance of the population can be small enough to make it possible to linearize. With increasing search space dimensionality, the error introduced by the linearization tends to zero, and even for moderate search space dimensionalities the predictions that can be obtained using the results from Section 2 describe the behavior of the (μ, λ)-ES on the noisy sphere with high accuracy. The close relationship of the sphere model to other models of optimization problems such as the parabolic ridge studied by Oyman et al. [12] makes it seem likely that the results that can be obtained with the approach pursued in the present paper can be used in future analyses of other fitness environments.

The objective function

$$f : \mathbb{R}^N \to \mathbb{R}$$
$$f(\mathbf{x}) = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}}(\mathbf{x} - \hat{\mathbf{x}}), \tag{12}$$

where the task is minimization and where $\hat{\mathbf{x}} \in \mathbb{R}^N$ denotes the optimizer, is commonly referred to as the sphere model and has been introduced by Rechenberg [14] as a model for unconstrained optimization problems at a stage where the population of candidate solutions is already in relatively close vicinity to the optimizer. It has been proven to be of great use for studying the scaling behavior of ES with respect to parameters such as the search space dimensionality N, the population size parameters μ and λ, and the mutation strength. In this section, as in Arnold and Beyer [3], we focus on the sphere model with Gaussian fitness-proportionate noise. That is, we assume that the noise level present when evaluating a candidate solution $\mathbf{x}$ is proportional to that candidate solution's ideal fitness. The noise strength $\sigma_\epsilon^* = \vartheta/(2f(\mathbf{x}))$ is independent of the location in search space. Fitness-proportionate noise models relative errors of measurement that arise for example in connection with physical measurement devices that are accurate up to a certain percentage of the quantity they measure.

The (μ, λ)-ES generates λ new candidate solutions from the parental population $\{\mathbf{x}_1, \ldots, \mathbf{x}_\mu\}$ by randomly picking one of the parents and adding a mutation vector the components of which are independently normally distributed with mean zero and with variance σ^2. We assume that some mechanism for adapting the mutation strength σ is in place, such as mutative self-adaptation (Rechenberg [14], Schwefel [16]). The distances from the optimizer are denoted as $R_i = \|\mathbf{x}_i - \hat{\mathbf{x}}\|$, and the average distance of the population from the optimizer is $R = \sum_{i=1}^{\mu} R_i/\mu$. The progress rate φ is the expected value of the change in R from one time step to the next. As common in ES theory, we introduce the normalized progress rate $\varphi^* = \varphi N/R$ and the normalized mutation strength $\sigma^* = \sigma N/R$.

Analyses of the local performance of ES on the sphere rely on a decomposition of mutation vectors into two components: a central component in direction of the optimizer and a lateral component in the plane perpendicular to that direction. Both Rechenberg [15] and Beyer [8] have shown that in the limit of infinite search space dimensionality, the distance from the optimizer $r = \|\mathbf{y} - \hat{\mathbf{x}}\|$ of an offspring $\mathbf{y}$ generated from parent $\mathbf{x}_i$ is

$$r = R_i - \sigma z + \frac{N\sigma^2}{2R_i},$$

where z is a standard normally distributed random variable. The second term on the right hand side is a result of the central component of the mutation vector, the third term is contributed by the lateral component. By normalization, it follows that

$$N\frac{R_i - r}{R} = \sigma^* z - \frac{R}{R_i}\frac{\sigma^{*2}}{2}. \tag{13}$$

Let us for the moment assume that the variance of the R_i decreases as the search space dimensionality increases. In that case, the second summand on the right hand side of Eq. (13) tends to $-\sigma^{*2}/2$ and is thus independent of the parent that the offspring candidate solution is generated from. Hence, by a simple linear transformation, we have the same situation as in Section 2, with the z-values of the offspring candidate solutions taking the role of the x_i in Section 2. The expected average of the selected z-values would thus be $c_{\mu,\lambda}(\vartheta)$, where $\vartheta = \sigma_\epsilon^*/\sigma^*$. Therefore, the progress rate of the (μ, λ)-ES on the noisy sphere in the limit of infinite search space dimensionality would be

$$\varphi^* = \sigma^* c_{\mu,\lambda}(\vartheta) - \frac{\sigma^{*2}}{2}. \tag{14}$$

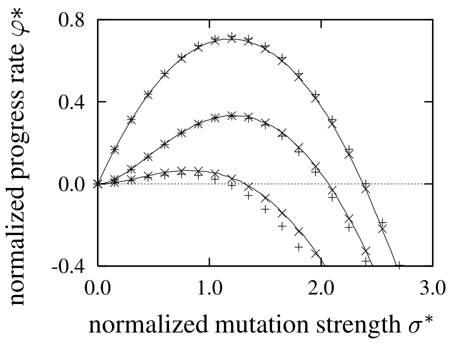

Figure 3: Normalized progress rate φ^* on the noisy sphere of a $(3, 10)$-ES as a function of normalized mutation strength σ^* for, from top to bottom, noise strengths $\sigma^*_\epsilon = 0.0$, 2.0, and 4.0. The solid lines mark results from Eq. (14). The crosses represent data from runs of evolution strategies with $N = 40$ $(+)$ and $N = 400$ $(\times)$.

For zero noise strength, Eq. (14) formally agrees with a progress rate law given by both Rechenberg [15] and Beyer [8]. Rechenberg also claimed the validity of the law in the presence of noise.

What remains to be seen is whether the variance of the R_i really decreases with increasing search space dimensionality. It is reasonable to assume — and can indeed be observed in experiments — that the population variance is largest if selection is random. Notice that this fact is reflected in the lower graph of Figure 1. The dynamics of EAs with random selection have been studied by Beyer [7]. In that reference, it is shown that the population variance of an ES with mutation strength σ and with random selection does not exceed $\mu\sigma^2$. Thus, the quotient $d_i = (R_i - R)/(\sqrt{\mu}\sigma)$ is of order unity in N. As

$$\frac{R_i}{R} = 1 + \sqrt{\mu}\frac{\sigma}{R}d_i = 1 + \sqrt{\mu}\frac{\sigma^*}{N}d_i,$$

and as for μ and σ^* fixed the last term tends to zero as N tends to infinity, it follows that R_i/R more and more closely approaches unity as the search space dimensionality increases. For sufficiently high search space dimensionality and not too large a population, Eq. (14) therefore indeed describes the progress rate of the (μ, λ)-ES on the noisy sphere.

Figure 3 compares predictions from Eq. (14) with empirical measurements of the progress rate of a $(3, 10)$-ES on the noisy sphere with search space dimensionalities $N = 40$ and $N = 400$. It can be seen that for $N = 40$ the agreement is quite good, but that the accuracy of the predictions afforded by Eq. (14) somewhat decreases with increasing noise strength and with increasing mutation strength. This is reasonable as both increasing noise strength and increasing mutation strength increase the population variance. Search space dimensionality $N = 400$ is sufficient for achieving very good agreement with empirical measurements across the entire range of noise and mutation strengths considered.

Using Eq. (14) it is now possible to determine optimal population sizes and maximal efficiencies on the noisy sphere for sufficiently high search space dimensionality. The efficiency η of the (μ, λ)-ES is defined as the normalized progress rate per evaluation of the objective function and thus as

$$\eta = \frac{\varphi^*}{\lambda}.$$

The division by λ serves the purpose of accounting for the computational costs that are assumed to be dominated by the cost of evaluating the fitness function. We determine numerically mutation strengths and population size parameters μ and λ that maximize the efficiency. As the focus in this section is the relevance of the linear progress law for the sphere, empirical values for the coefficients $c_{\mu,\lambda}(\vartheta)$ have been used so as to not introduce errors resulting from the approximation made in Section 2. While for low noise strengths ($0.0 \leq \sigma^*_\epsilon \leq 4.0$) there are virtually no discrepancies between the resulting curves, for higher noise strengths the dependence of the progress rate on the population size parameters becomes so weak that some deviations in the results can be observed.

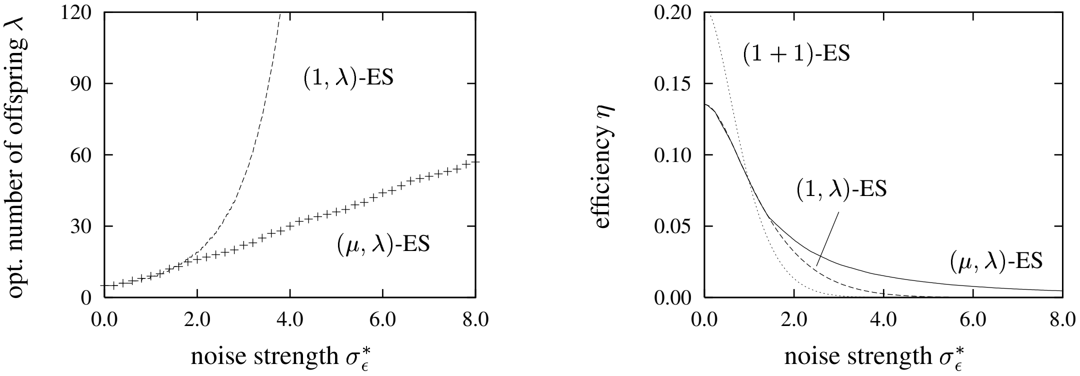

Figure 4: Optimal number of offspring per time step λ and maximal efficiency η on a high-dimensional sphere as functions of noise strength σ_ϵ^*. In the left hand graph, the dashed line represents results for the $(1, \lambda)$-ES, the crosses for the (μ, λ)-ES with optimally chosen μ. In the right hand graph, the solid and dashed lines correspond to the (μ, λ)-ES and the $(1, \lambda)$-ES, respectively, with optimally chosen population size parameters. The dotted line reflects the result for the $(1 + 1)$-ES that was obtained by Arnold and Beyer [3].

The left hand graph of Figure 4 shows the optimal number of offspring λ as a function of the noise strength for the (μ, λ)-ES with optimally chosen μ as well as for the $(1, \lambda)$-ES. It can be seen that, except for very low noise strengths, the (μ, λ)-ES ideally operates with many fewer offspring candidate solutions per time step than the $(1, \lambda)$-ES. At least for the range considered, the relationship between the noise strength and the optimal number of offspring candidate solutions of the (μ, λ)-ES appears to be nearly linear. The optimal truncation ratio μ/λ consistently lies between 0.1 and about one quarter, with a tendency to ratios at the upper end of that interval as the noise strength increases.

The right hand graph of Figure 4 compares the efficiency of the (μ, λ)-ES with optimally chosen population size parameters μ and λ with those of the $(1, \lambda)$-ES with optimally chosen λ and of the $(1 + 1)$-ES. The efficiency of the $(1 + 1)$-ES exceeds the efficiency of the other two strategies only up to a noise strength of $\sigma_\epsilon^* \approx 1.0$ and is markedly inferior for higher noise strengths. Up to a noise strength of about $\sigma_\epsilon^* \approx 1.4$ it is not useful to retain more than a single candidate solution and the curves for the (μ, λ)-ES and the $(1, \lambda)$-ES agree. Above this noise strength the efficiency of the (μ, λ)-ES can significantly exceed that of the $(1, \lambda)$-ES. It is important to note that the sensitivity of the efficiency of the (μ, λ)-ES to the population size parameters μ and λ is low especially for high noise strengths. The left hand graph of Figure 4 in combination with the observation on optimal truncation ratios made above can often serve as a guideline for choosing population size parameters that result in near-optimal performance.

5 Conclusions

In this paper, the influence of distributed populations on the performance of ES in continuous search spaces has been studied. In particular, the behavior of the (μ, λ)-ES on a linear fitness function was analyzed using a moment-based approach for describing the population of candidate solutions. The results of the analysis that considered moments up to the fourth order and that neglected fluctuations proved to yield good estimates for the population variance as well as for the expected progress. Based on those results, we identified the population variance as a quantity of great importance for the understanding of the (μ, λ)-ES. We have seen that it contributes to the "signal strength" of the selection process, and that greater variances improve the signal-to-noise ratio that the strategy operates under. However, as greater variances can be achieved only by increasing the proportion of candidate solutions that are retained and thus at the price of reduced selection pressure, there exists an optimal truncation ratio that depends on the population size as well as on the noise strength. We have seen that the optimal truncation ratio increases as the noise strength increases, but that attaining it exactly becomes less significant as the dependence of the quality gain on it becomes weaker.

The results obtained on the linear fitness function are of immediate relevance and of practical interest whenever it is possible to linearize the fitness function at hand. By considering the sphere model, we have seen that in high-dimensional search spaces such a linearization can indeed be possible. Determining optimal population sizes on the sphere we found that the optimal number of offspring candidate solutions generated per time step is much lower for the (μ, λ)-ES than for the $(1, \lambda)$-ES — a result of great interest to the ES practitioner —, and that above a certain noise strength substantial performance gains can be achieved by retaining more than the (seemingly) best candidate solution. Moreover, it has been seen that the choice of population size parameters is relatively uncritical, and that retaining about 20% of the total number of candidate solutions generated per time step almost universally yields near-optimal performance. There is substantial hope that the results obtained in this paper are of relevance for future studies of the effects of noise in other high-dimensional fitness environments.

Appendix: Mathematica Program

This appendix contains the *Mathematica* program used to determine the population moments that result from the analysis in Section 2 and that have been used to generate Figure 1. The code up to and including the definition of `MakeSum` is a duplication of that proposed by Arnold and Beyer [4] for the computation of the expected values of the S_A. The remainder of the code determines numerically the coefficients $h^{i,k}_{\mu,\lambda}$ defined in Eq. (8) and solves the system of Eqs. (10).

```
Hermite[k_, x_] := Simplify[HermiteH[k, x/Sqrt[2]]/Sqrt[2]^k];

MakePi[A_] :=
  Apply[Plus,
    Map[Apply[Times, MapIndexed[x[First[#2]]^#1 &, #1]] &,
        Permutations[A]]];

MakeIntegrand1[A_] :=
  MakePi[A]*Product[1
                  + g1*Hermite[3, x[i]]
                  + g2*Hermite[4, x[i]]
                  + g1^2*Hermite[6, x[i]]/2,
                  {i, 1, Length[A]}];

HermiteExpand[expr_, x_] :=
  ToHermite[Expand[expr]
    /. {(g1^i_ /; i>2)->0, g1^i_.*g2^j_.->0, g2^i_->0}, x];

ToHermite[expr1_+expr2_, x_] :=
```

```
  ToHermite[expr1, x]+ToHermite[expr2, x];
ToHermite[expr_, x_] :=
  expr*He[0, x] /; Not[MatchQ[expr, a_.*x^k_.]];
ToHermite[expr_. x_^k_., x_] :=
  expr*He[k, x]
    + ToHermite[Expand[expr*(x^k-Hermite[k, x])], x];

Integrate1[A_, 0] := A;
Integrate1[A_, i_] :=
  Integrate1[Int1[HermiteExpand[A, x[i]], x[i], y[i]], i-1];

Int1[c_ expr_, x_, y_] := c Int1[expr, x, y] /; FreeQ[c, x];
Int1[expr1_+expr2_, x_, y_] :=
  Int1[expr1, x, y]+Int1[expr2, x, y];
Int1[He[k_, x_], x_, y_] := a^(k+1) Hermite[k, y] g[y];

MakeIntegrand2[A_] := Integrate1[MakeIntegrand1[A], Length[A]];

Integrate2[A_, 0] := A;
Integrate2[A_, i_] :=
  Integrate2[Int2[HermiteExpand[A, y[i]], y[i], y[i-1]], i-1];

Int2[c_ expr_, x_, y_] := c Int2[expr, x, y] /; FreeQ[c, x];
Int2[expr1_+expr2_, x_, y_] :=
  Int2[expr1, x, y]+Int2[expr2, x, y];

Int2[He[0, x_] g[x_], x_, y_] := f[y];
Int2[He[0, x_] f[x_] g[x_], x_, y_]:= f[y]^2/2;
Int2[He[1, x_] g[x_]^b_., x_, y_] := g[y]^b/b;
Int2[He[1, x_] f[x_] g[x_]^2, x_, y_] :=
  f[y]g[y]^2/2-Int2[He[0, x] g[x]^3, x, y]/2;
Int2[He[k_, x_] g[x_], x_, y_] := Hermite[k-1, y]g[y];
Int2[He[k_, x_] g[x_]^b_., x_, y_] :=
  (Hermite[k-1, y]g[y]^b/b
    - (b-1)(k-1)Int2[He[k-2, x] g[x]^b, x, y]/b) /; k>=2;
Int2[He[k_, x_] f[x_] g[x_], x_, y_] :=
  (Hermite[k-1, y]f[y]g[y]
    - Int2[He[k-1, x] g[x]^2, x, y]) /; k>=1;
Int2[He[k_, x_] f[x_] g[x_]^2, x_, y_] :=
  (Hermite[k-1, y]f[y]g[y]^2/2
    - (k-1)Int2[He[k-2, x] f[x] g[x]^2, x, y]/2
    - Int2[He[k-1, x] g[x]^3, x, y]/2) /; k>=2;

Substitution[c_] := c /; FreeQ[c, y[0]];
Substitution[expr1_+expr2_] :=
  Substitution[expr1] + Substitution[expr2];
Substitution[expr1_*expr2_] :=
```

```
  Substitution[expr1] * Substitution[expr2];
d1 = g1 a^3(x^2-1)+g2 a^4(x^3-3x)-g1^2 a^6(2x^3-5x);
d2 = g1^2 a^6(x^2-1)^2;

Substitution[y[0]^k_.] :=
  x^k + k x^(k-1)d1 + k(k-1)x^(k-2)d2/2;
Substitution[f[y[0]]^k_.] :=
  f[x]^k - k f[x]^(k-1)g[x](d1-x*d2/2)
    + k(k-1)f[x]^(k-2)g[x]^2 d2/2;
Substitution[g[y[0]]^k_.] :=
  g[x]^k(1 - k(x*d1-(x^2-1)d2/2) + k(k-1)x^2d2/2);

MakeSum[A_] :=
  HermiteExpand[
    Substitution[Integrate2[MakeIntegrand2[A], Length[A]]]
      /a^Length[A], x];

$RecursionLimit=512

h[i_, k_, mu_, lambda_] := h[i, k, mu, lambda] =
  N[(lambda-mu)Binomial[lambda, mu]Integrate[
      Hermite[k, x]
      (Exp[-x^2/2]/Sqrt[2Pi])^(i+1)
      ((1+Erf[x/Sqrt[2]])/2)^(lambda-mu-1)
      ((1-Erf[x/Sqrt[2]])/2)^(mu-i),
    {x, -Infinity, Infinity}]];

DetermineMoments[lambda_, theta_] :=
(
  Clear[a];

  NEval[arg_] :=
    arg //. { g[x]^i_.He[k_, x]:>h[i, k, mu, lambda],
               He[k_, x]:>h[0, k, mu, lambda],
               f[x]^j_.:>1,
               a:>Sqrt[(1+g0)/(1+theta^2+g0)] };

  For[mu=1, mu<=lambda, mu+=1,
    s1 = NEval[MakeSum[{1}]];
    s11 = If[mu<=1, 0, NEval[MakeSum[{1, 1}]]];
    s111 = If[mu<=2, 0, NEval[MakeSum[{1, 1, 1}]]];
    s1111 = If[mu<=3, 0, NEval[MakeSum[{1, 1, 1, 1}]]];
    s2 = NEval[MakeSum[{2}]];
    s21 = If[mu<=1, 0, NEval[MakeSum[{2, 1}]]];
    s211 = If[mu<=2, 0, NEval[MakeSum[{2, 1, 1}]]];
    s22 = If[mu<=1, 0, NEval[MakeSum[{2, 2}]]];
    s3 = NEval[MakeSum[{3}]];
```

```
    s31 = If[mu<=1, 0, NEval[MakeSum[{3, 1}]]];
    s4 = NEval[MakeSum[{4}]];

    A1 = s1;
    A2 = (mu-1)(s2-2s11)/mu;
    A3 = (mu-1)(mu-2)(s3-3s21+12s111)/mu^2;
    A4 = (mu-1)(mu^2-6mu+6)(s4-4s31+6s22)/mu^3
           -12(mu-1)(mu-2)(mu-3)(s22-2s211+12s1111)/mu^3;

    rule = FindRoot[{g0==(1+g0)A2, 6 g1==A3, 24 g2==A4},
                    {g0, 1}, {g1, 0}, {g2, 0}];

    Print[mu, " ", Sqrt[1+g0]A1 /. rule, " ", g0 /. rule];
  ];
);
```


Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants Be1578/4-2 and Be1578/6-3. The publication of the work was also supported by the DFG as part of the Collaborative Research Center "Computational Intelligence" (SFB 531). Hans-Georg Beyer is a Heisenberg Fellow of the DFG.

References

[1] Arnold, D. V. (2002). *Noisy Optimization with Evolution Strategies*, Kluwer Academic Publishers, Norwell, Massachusetts.

[2] Arnold, D. V. and Beyer, H.-G. (2001). Investigation of the (μ, λ)-ES in the presence of noise. Proceedings of the 2001 IEEE Congress on Evolutionary Computation. Seoul, Korea, May 27–30, pages 332–339.

[3] Arnold, D. V. and Beyer, H.-G. (2002). Local performance of the $(1 + 1)$-ES in a noisy environment. *IEEE Transactions on Evolutionary Computation*, 6(1):30–41.

[4] Arnold, D. V. and Beyer, H.-G. (2002). Expected sample moments of concomitants of selected order statistics. Technical Report CI 134/02, SFB 531, University of Dortmund, Dortmund, Germany.

[5] Beyer, H.-G. (1993). Toward a theory of evolution strategies: Some asymptotical results from the $(1 \overset{+}{,} \lambda)$-theory. *Evolutionary Computation*, 1(2):165-188.

[6] Beyer, H.-G. (1995). Toward a theory of evolution strategies: The (μ, λ)-theory. *Evolutionary Computation*, 2(4):381–407.

[7] Beyer, H.-G. (1999). On the dynamics of EAs without selection. In Banzhaf, W. and Reeves, C., editors, *Foundations of Genetic Algorithms 5*, pages 5–26, Morgan Kaufmann, San Mateo, California.

[8] Beyer, H.-G. (2001). *The Theory of Evolution Strategies*, Springer Verlag, Berlin, Germany.

[9] David, H. A. and Nagaraja, H. N. (1998). Concomitants of order statistics. In Balakrishnan, N. and Rao, C. R., editors, *Handbook of Order Statistics 16*, pages 487–513, Elsevier, Amsterdam, Netherlands.

[10] Miller, B. L. and Goldberg, D. E. (1997). Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 4(2):113–131.

[11] Nissen, V. and Propach, J. (1998). Optimization with noisy function evaluations. In Eiben, A. E. et al., editors, *Parallel Problem Solving from Nature V*, pages 159–168, Springer Verlag, Berlin, Germany.

[12] Oyman, A. I., Beyer, H.-G., and Schwefel, H.-P. (2000). Analysis of a simple ES on the "parabolic ridge". *Evolutionary Computation*, 8(3):249–265.

[13] Rattray, M. and Shapiro, J. L. (1997). Noisy fitness evaluation in genetic algorithms and the dynamics of learning. In Belew, R. K. and Vose, M. D., editors, *Foundations of Genetic Algorithms 4*, pages 117–139, Morgan Kaufmann, San Mateo, California.

[14] Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach den Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart, Germany.

[15] Rechenberg, I. (1994). *Evolutionsstrategie '94*, Frommann-Holzboog, Stuttgart, Germany.

[16] Schwefel, H.-P. (1995). *Evolution and Optimum Seeking*, Wiley, New York, New York.

[17] Stuart, A. and Ord, J. K. (1994). *Kendall's Advanced Theory of Statistics*, Sixth edition, Volume I: Distribution Theory, Arnold, London, United Kingdom.