The influence of stochastic quality functions on evolutionary search

Bernhard Sendhoff, Hans-Georg Beyer, Markus Olhofer

2004

Preprint:

In "Recent Advances in Simulated Evolution and Learning" K.C. Tan, M.H. Lim, X. Yao and L. Wang (Eds.), World Scientific , 2004.

CHAPTER 9

THE INFLUENCE OF STOCHASTIC QUALITY FUNCTIONS ON EVOLUTIONARY SEARCH

Bernhard Sendhoff[1], Hans-Georg Beyer[2] and Markus Olhofer[1]

[1] *Honda Research Institute Europe GmbH,*
Carl-Legien-Str. 30, 63073 Offenbach, Germany
[2] *Dept. of Computer Science XI, University of Dortmund,*
44221 Dortmund, Germany

In this chapter, we will analyse the influence of noise on the search behaviour of evolutionary algorithms. We will introduce different classes of functions which go beyond the simple additive noise model. The first function demonstrates a trade-off between an expectation and a variance based measure for the evaluation of the quality in the context of stochastic optimisation problems. Thereafter, we concentrate on functions whose topology is changed when the expectation value is taken as the quality criterion. In particular, for functions with noise induced multi-modality (FNIM), the process can be regarded as a bifurcation. The behaviour of two types of evolution strategies is analysed for FNIMs.

1. Introduction

Optimisation in the presence of noise has been studied as early as 1970 in the area of stochastic programming[7]. In stochastic programming the objective function and possibly the constraints are subject to stochastic perturbations. The standard approach for these cases is to work on the expectation value of the objective functions and therefore, to render the optimisation problem deterministic. The remaining problem is that the evaluation of the expectation value might involve a prohibitively large number of function evaluations. Therefore, one is left to estimate the expectation value with a residual error. Since evolutionary algorithms are believed to be particularly robust optimisation techniques, their application to noisy objective functions seems particularly suitable. Fitzpatrick and Grefenstette[9] and later Aizawa and Wah[1] analysed genetic algorithms in a noisy environment. The influence of noise on the performance of evolution strategies was first dis-

cussed by Beyer[5]. Tsutsui and Gosh[13] claimed that it is not necessary to explicitly calculate the expectation value for each solution but instead that it is sufficient to evaluate the solution once and that the population inherently ascertains that the expectation value of the objective function is the target of the optimisation. Although their analysis relies on the schema theorem and is therefore restricted to proportional selection, empirical results show that indeed evaluating each individual solution only once instead of estimating the expectation value of the objective function can be most efficient.

In almost all cases of practical relevance, it is impossible to evaluate the expectation value analytically, therefore, it has to be estimated. Whether we estimate it explicitly by using a sample of size K or implicitly by exploiting the population, the optimisation method has to cope with statistical fluctuations. Therefore, although theoretically possible, practically, the stochastic problem can never be reduced to a deterministic one.

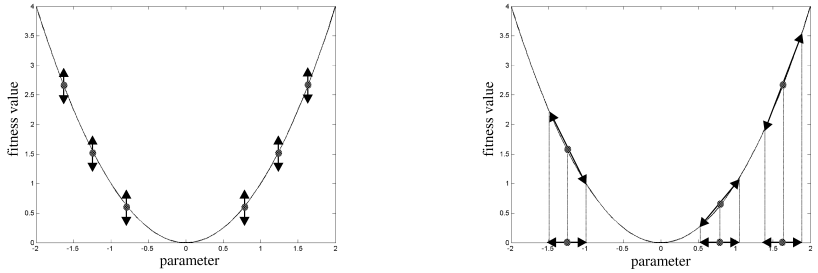
A stochastic problem is one where the quality landscape that the optimisation algorithms "sees" differs non-deterministically between two evaluations. The character of this difference depends on three main aspects:

The type of noise. Most analytical results on noisy evolutionary optimisation were obtained for the simple additive noise model, i.e., the noise term (usually normally or uniformly distributed) is added to the objective function value[2]. This case is depicted in Figure 1(a). Although interesting analytical results for the algorithm, e.g. on the role of the population, have been discovered, the character of the quality landscape is not changed. This differs for systematic noise models, where the noise term is not restricted to be additive but can occur anywhere inside the quality function. Robustness constraints on the optimisation, see e.g. Wiesmann *et al.*[14] and Branke[8], constitute a special case of the systematic noise models, where the noise term is added to the parameter set, which is depicted in Figure 1(b). In this chapter, we will concentrate on the systematic noise model.

The quality function. Every optimisation problem is unique depending on the quality function. However, in order to be able to apply empirical or analytical results to a class of problems, test functions are devised which are simple enough to be susceptible for analysis while capturing the specificity of the problem class. Examples are the sphere, the ridge function or the Ackley function for multi-

modal search landscapes. In this chapter, we will propose and analyse a class of test functions which has been called functions with *noise-induced multi-modality (NIMM)* by Sendhoff *et al.*[12].

The evaluation criterion. For deterministic single-criteria optimisation problems the evaluation criterion is usually the minimisation or the maximisation of the quality function possibly according to some constraints. For noisy optimisation problems the expectation value is frequently used as the evaluation criterion although other choices are possible. Beyer *et al.*[6] proposed a differentiation between statistical momentum based criteria and threshold criteria, where (for maximisation problems) the probability that the quality value is below a certain threshold is minimised. It should be noted that the choice of the evaluation criterion can fundamentally alter the characteristics of the search landscape. In the next sections, we will apply the expectation value as a evaluation criterion mostly because the analysis involved in the threshold criterion is slightly trickier. However, we will also use the variance as an additional measure in Section 2.



(a) (b)

Fig. 1. (a) Variations due to additive noise on the quality function. The dots represent the fitnesses of individual solutions and the arrows the direction of variation, the underlying curve defines a one-dimensional fitness landscape. (b) Variations due to noise on the parameters resulting in variations along the fitness curve.

Stochastic problems are always (at least practically) dynamic optimisation problems[8]. However, in this chapter we will call problems *stochastic* if the time-scale of their change is fast compared to the change of the individual solutions and *dynamic* if this time-scale is slow compared to the change of the individual solutions. The latter problem is not the subject of this

work. After the introduction of different noise models in the next section, we will analyse the behaviour of two types of evolution strategies for the new noise model, the NIM functions. In Section 4, we will extend the FNIM to higher dimensions. In Section 5, we will summarise this chapter.

2. Classes of Noisy Optimization Problems

As we pointed out in the introduction, the character of a noisy optimisation problem[a] is mainly determined by three factors. The type of noise, the quality function and the evaluation criterion. In this section, we will introduce three qualitatively different cases of noisy optimisation problems, which display a particular characteristic property. The first extension of additive noise models is the sphere function with a noise term added to the objective parameters $F(\mathbf{x}) = (\mathbf{x} + \mathbf{z})^2$, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \varepsilon^2\mathbf{1})$. Here $\mathcal{N}(\mathbf{0}, \varepsilon^2\mathbf{1})$ denotes a vector of random numbers, where each component is normally distributed with zero mean and variance ε^2. Beyer *et al.*[6] have shown that for this function the minimum of the original sphere model ($\mathbf{x} = \mathbf{0}$) coincides with the minimum of the expectation value and of the variance of the quality function with systematic noise. Although further analysis nevertheless reveals some interesting properties in particular for the threshold measure, we will not discuss this case any further.

2.1. *Expectation Value – Variance Trade-Off*

Although usually only the expectation value is used for noisy optimisation problems, in particular for robust solutions it is often the minimisation of the variance which is needed for practical applications. As we noted above for the sphere model with systematic noise, minimisation of the expectation value also leads to minimisation of the variance. However, the following function shows that this does not hold in general:

$$F_1(\mathbf{x}) = (\mathbf{x}^2 - 1)\, z + \mathbf{x}^2, \quad z \sim \mathcal{N}(0, \varepsilon^2), \mathbf{x} \in I\!R^N. \tag{1}$$

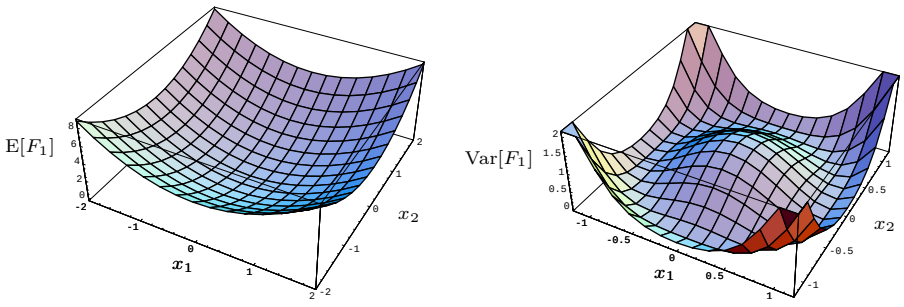
Here $\mathcal{N}(0, \varepsilon^2)$ denotes a Gaussian distributed random number with zero mean and variance ε^2. The calculation of the expectation value and the variance gives:

$$\mathrm{E}[F_1(\mathbf{x})|\mathbf{x}] = \mathbf{x}^2, \tag{2}$$

$$\mathrm{Var}[F_1(\mathbf{x})|\mathbf{x}] = \varepsilon^2\,(\mathbf{x}^2 - 1)^2. \tag{3}$$

[a]Note, that in this chapter we will generally deal with minimisation problems unless otherwise stated.

Therefore, the minima of the expectation value and the variance are given by $\mathbf{x} = \mathbf{0}$ and $\mathbf{x}^2 = 1$. The minimum of the expectation value corresponds to a local maximum of the variance, as shown in Figs. 2(a) and (b). For func-



(a) (b)

Fig. 2. The expectation value (a) and the variance (b) of function F_1 for $\varepsilon^2 = 1$, $N = 2$.

tions like F_1 the expectation value and the variance cannot be minimised at the same time and the problem basically constitutes a multi-objective optimisation problem, see Jin and Sendhoff [11]. We note, that the characteristic of function F_1 is that the noise term z is multiplied both to the parameter values $\mathbf{x}$ as well as to a constant term (in this example "1"). If we replace the constant term by an external parameter, say a, this means that stochastic variations of the parameter values and of an external parameter, e.g. the cruising speed for an aerodynamic optimisation problem, share the same source which is not unlikely to occur in certain applications.

2.2. *Topological Changes of the Quality Landscape*

If the number of optima of the expectation value of functions is different from the noise-free case, we call these changes *topological*. These functions do not have to be very complex, like previously proposed in the literature[8,14], all that is needed is a noise induced variation between two minima. As one can easily imagine averaging over such functions will (in some cases) result in merging minima and thereby absorbing and erasing the maxima in between the minima:

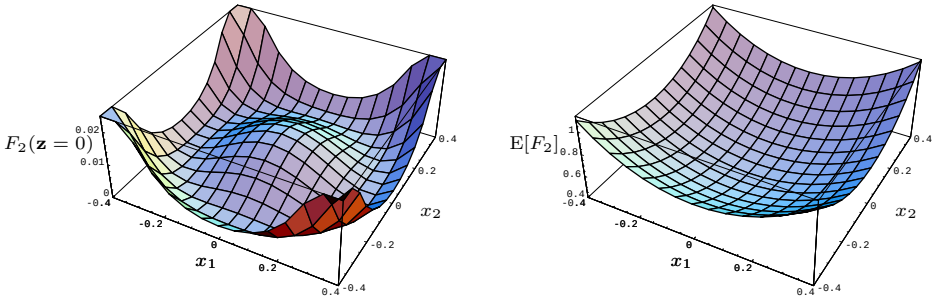
$$F_2(\mathbf{x}) = \left((\mathbf{x}+\mathbf{z})^2 - a\right)^2, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \varepsilon^2\mathbf{1}), \mathbf{x} \in I\!R^N. \tag{4}$$

Here $\mathcal{N}(\mathbf{0}, \varepsilon^2\mathbf{1})$ denotes a vector of random numbers, where each component is normally distributed with zero mean and variance ε^2. Using $\mathrm{E}[z^2] = \varepsilon^2$,

$\mathrm{E}[z^3] = 0$ and $\mathrm{E}[z^4] = 3\varepsilon^4$, for $z \sim \mathcal{N}(0, \varepsilon^2)$, the calculation of the expectation value gives:

$$\mathrm{E}[F_2(\mathbf{x})|\mathbf{x}] = (\mathbf{x}^2)^2 + 2\mathbf{x}^2 \left((N+2)\varepsilon^2 - a\right) + \varepsilon^2 \left(N(N+2)\varepsilon^2 - 2aN\right) + a^2. \quad (5)$$

In Figure 3 function F_2 without noise ($\mathbf{z} = 0$) is shown together with the expectation value $\mathrm{E}[F_2|\mathbf{x}]$. The minima are merged into one global minimum at $(0, 0)$ replacing the local maximum. A closer look at Eq. (5) reveals



(a) (b)

Fig. 3. Figure (a): function F_2 without noise for $N = 2$ and $a = 0.1$; figure (b): expectation value of F_2 for $N = 2$ and $a = 0.1$.

that this transition depends on the variance of the noise ε^2 and on the distance between the two minima of the noise free function controlled by the parameter a. The transition occurs if

$$\varepsilon > \left(\frac{a}{N+2}\right)^{\frac{1}{2}}. \quad (6)$$

The dependence on a is shown in Figure 4. Figure 3(a) also shows that function F_2 without noise does not have one optimum but infinitely many optima. If we compare this subspace of optimal solutions to the Pareto space in multi-objective optimisation where we also encounter a possibly continuous set of solutions of identical quality, it is reasonable to say that the identification of the whole space should be the target of the optimisation. We will come back to this point in Section 5. Here we only note that the optimal manifold of function F_2 without noise is given by the hyper-sphere

$$\mathbf{x}^2 = a. \quad (7)$$

For $N = 2$ the manifold is a circle with diameter a as shown in Figure 3(a).

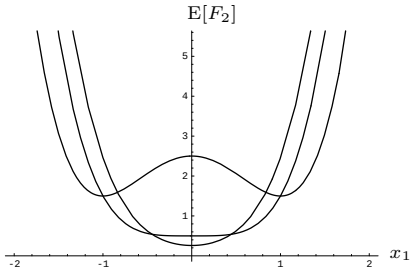


Fig. 4. The expectation value of function F_2 for $\varepsilon^2 = 0.25$, $x_2 = 0$, $N = 2$ and $a = 0.4, 1.0, 2.0$ (from top to bottom). As predicted by equation (6), the transition occurs below the critical value of $a = 1$.

2.3. *Functions with Noise Induced Multi-Modality (FNIM)*

The transition from single-modal to multi-modal characteristics in a class of functions under the influence of noise is less straightforward than the merging of optima demonstrated in the last section for function F_2.

The introduction of the FNIM in this section is motivated by the qualitative behaviour of evolution strategies for the design optimisation of gas-turbine blades. The behaviour suggests that the local fitness space might look similar to the fitness function shown in Figure 5. Since the dimensionality of the parameter space of the design optimisation problem is much higher, the model can only be regarded as one possible interpretation. In

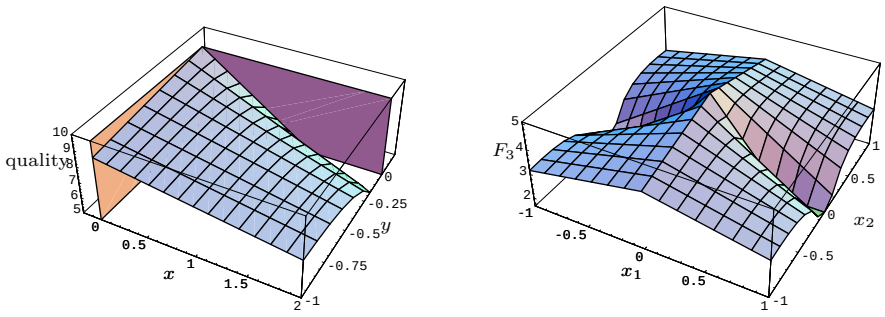


Fig. 5. Left: Qualitative model for the local fitness landscape motivated by the behaviour of evolution strategies for the design optimisation of gas-turbine blades. Right: Function F_3 with $n = 2, z = 0, a = 5$ and $b = 0.2$.

the direction of the y-axis (assuming $x = 0$) the fitness increases nearly linear along a ridge. The downwards slope from the ridge in the positive x-direction significantly increases with increasing fitness value. The fitness

space is bounded by two regions of infeasible solutions (shown by the filled rectangles) for example due to geometric constraints or unstable results of the fluid-dynamics flow solver. Needless to say that the position of the infeasible region does not exactly coincide with the ridge, but can lie somewhere on the negative x-axis. Noise is introduced in this fitness model by demanding robustness of the parameter representing the x-direction perpendicular to the ridge. Thus, the resulting design should display stable performance under variations of the x-parameter. The optimum of the $x-$averaged fitness landscape will not remain at the $(y = 0)-$boundary to the infeasible region but move along the ridge to smaller y-values assuming that the increase of the downwards slope in the x-direction is sufficient.

Function F_3, Eq. (8), displays the linear increase along the ridge and the sharp decrease in the x_{N-1}-coordinate in the vicinity of the optimum at $(0, 0)$.

$$F_3(\mathbf{x}) = a - \frac{|x_{N-1} + z| + \sum_{i=1}^{N-2} x_i^2}{|x_N| + b} - |x_N|$$
$$z \sim \mathcal{N}(0, \varepsilon^2),\ b > 0,\ \mathbf{x} \in I\!R^N. \tag{8}$$

In order to be able to neglect the infeasible regions in the analysis, function F_3 has been designed in such a way that a clear optimum exists when no robustness is taken into account. Thus, without noise ($z = 0$) F_3 is a uni-modal function, as shown in Figure 5 for $N = 2$. Next, we derive $\mathrm{E}[F_3|\mathbf{x}]$:

$$\mathrm{E}[F_3|\mathbf{x}] = a - \frac{E[|x_{N-1} + z|] + \sum_{i=1}^{N-2} x_i^2}{|x_N| + b} - |x_N|. \tag{9}$$

For $z \sim \mathcal{N}(0, \varepsilon^2)$, $\mathrm{E}[|x + z|]$ is given as follows:

$$\mathrm{E}[|x + z|] = \mathrm{E}[x + z]_{z>-x} + \mathrm{E}[-(x + z)]_{z<-x} \tag{10}$$
$$= \frac{1}{\sqrt{2\pi\varepsilon^2}} \left(x \int_{-x}^{x} e^{(\frac{-z^2}{2\varepsilon^2})} dz + 2 \int_{x}^{\infty} z\, e^{(\frac{-z^2}{2\varepsilon^2})} dz \right) := \xi(x) \tag{11}$$

Using (11), we get

$$\mathrm{E}[F_3|\mathbf{x}] = a - \frac{\xi(x_{N-1}) + \sum_{i=1}^{N-2} x_i^2}{|x_N| + b} - |x_N|. \tag{12}$$

$\mathrm{E}[F_3|\mathbf{x}]$ is shown in Figure 6 for fixed values of b and ε. In particular when we observe the 2D cross section shown in Figure 6, it is evident that the uni-modal function has changed into a bi-modal function due to averaging over the variations in one of the design parameters.

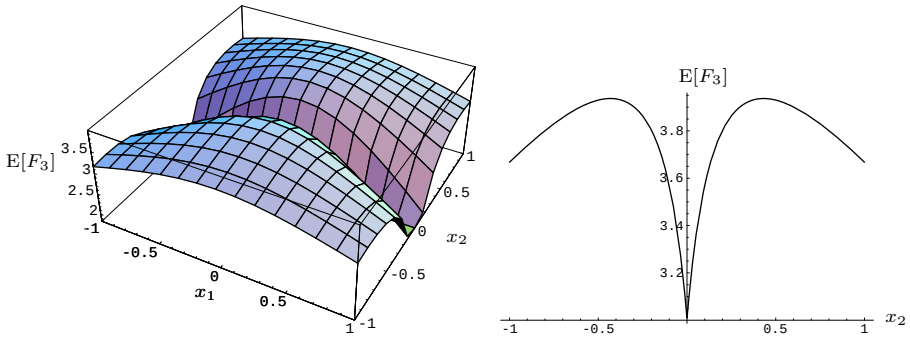


Fig. 6. Left: $\mathrm{E}[F_3|\mathbf{x}]$ with $N = 2$, $a = 5, b = 0.2$ and $\varepsilon^2 = 0.25$. Right: Two-dimensional cross section ($x_1 = 0$).

Qualitatively, this process of changing a uni-modal fitness function into a multi-modal function (or in our example into a bi-modal function) by averaging over the variations of one parameter is similar to a *bifurcation* process. The global maximum becomes a local minimum and two new local maxima (of the same height) occur. The bifurcation depends on the parameter b and on the noise strength, the variance ε^2. Numerically, this dependence is shown in Figure 7. We note that for large b values and for small variances no bifurcation occurs. Both dependencies are easily understood. The parameter b governs the steepness of the slope near the optimum $(0, 0)$; the smaller b, the steeper the slope. The noise strength determines the fluctuation along the coordinate x_{N-1}. Together they both determine whether the single optimum will persist or whether it will bifurcate.

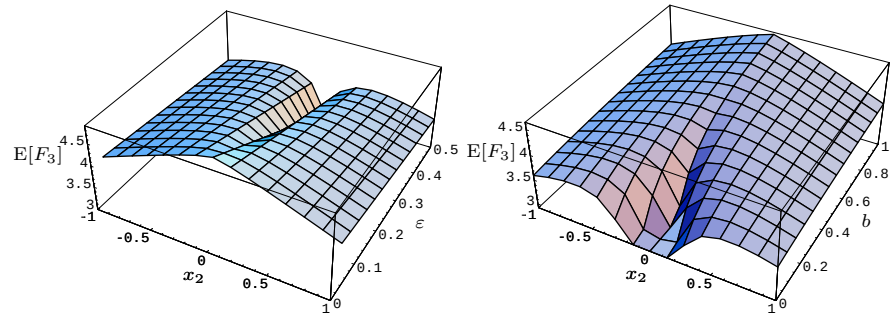


Fig. 7. The dependence of the bifurcation from a uni-modal to a bi-modal function on the standard deviation ε (left figure) and on the parameter b (right figure).

In order to analytically investigate the bifurcation behaviour further, Eq. (11) is too complex. Therefore, we smooth out the ridge and the slope in Eq. (8) and arrive at the following function F_4 which qualitatively shows a similar behaviour as F_3, as shown in Figure 8 for different levels of noise.

$$F_4(\mathbf{x}) = a - \frac{(x_{N-1} + z)^2 + \sum_{i=1}^{N-2} x_i^2}{x_N^2 + b} - x_N^2,$$
$$z \sim \mathcal{N}(0, \varepsilon^2),\ b > 0,\ \mathbf{x} \in I\!R^N. \tag{13}$$

Calculating the conditional expectation of F_4 is an easy task. Using $\mathrm{E}[(x_{N-1} + z)^2] = x_{N-1}^2 + \varepsilon^2$, we get:

$$\mathrm{E}[F_4|\mathbf{x}] = a - \frac{\varepsilon^2 + \sum_{i=1}^{N-1} x_i^2}{x_N^2 + b} - x_N^2. \tag{14}$$

Furthermore, it is straightforward to generalise F_4 and $\mathrm{E}[F_4|\mathbf{x}]$ to the multimodal case:

$$F_5(\mathbf{x}) = a - \frac{\sum_{i=1}^{N_1}(x_i + z_i)^2 + \sum_{i=N_2+1}^{N} x_i^2}{b + \sum_{N_1+1}^{N_2} x_i^2} - \sum_{N_1+1}^{N_2} x_i^2, \tag{15}$$

$$z_i \sim \mathcal{N}(0, \varepsilon^2),\ b > 0,\ \mathbf{x} \in I\!R^N,\ N_1 < N_2 \leq N. \tag{16}$$

Function F_4 is a special case of F_5 with $N_1 = 1, N_2 = 2$ (note that the indices are changed). We will come back to function F_5 in Section 5.

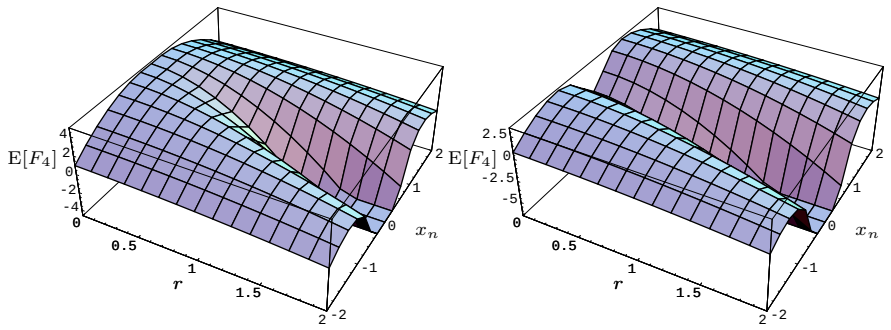


Fig. 8. Expected value landscapes of F_4 given by Eq. (14) for $\varepsilon = 0.25$ (left) and $\varepsilon = 1.0$ (right) - $(a = 5, b = 0.2)$.

Writing $\sqrt{\sum_{i=1}^{N-1} x_i^2} := r$, we can further shorten $\mathrm{E}[F_4|\mathbf{x}]$:

$$\mathrm{E}[F_4|\mathbf{x}] = a - \frac{r^2 + \varepsilon^2}{x_N^2 + b} - x_N^2. \tag{17}$$

The conditional variance is given by

$$\mathrm{Var}[F_4|\mathbf{x}] = \frac{4\varepsilon^2}{(x_N^2+b)^2}(x_{N-1}^2+\varepsilon^2/2). \tag{18}$$

Now we are in a position to determine the extrema of function F_4 by taking the partial derivative of (17) with respect to x_N and setting it to zero

$$\frac{\partial \mathrm{E}[F_4|\mathbf{x}]}{\partial x_N} = \frac{2x_N(r^2+\varepsilon^2)}{(x_N^2+b)^2} - 2x_N \stackrel{!}{=} 0. \tag{19}$$

Solving for x_N one gets the $\hat{x}_N$ points of the local optima. Besides the trivial solution $\hat{x}_N = 0$ there exist also nontrivial ones:

$$\hat{x}_N = \pm\sqrt{\sqrt{r^2+\varepsilon^2}-b} \quad \text{for} \quad r > \sqrt{b^2-\varepsilon^2}. \tag{20}$$

A closer examination of $\mathrm{E}[F_4|\mathbf{x}]$ reveals that there is a single maximum as long as the square root on the left-hand side in Eq. (20) is imaginary, i.e., for $\sqrt{r^2+\varepsilon^2}-b<0$. In this case the maximum is located at $(r, x_N) = (0,0)$ and the maximality condition for $\hat{x}_N = 0$ becomes $\varepsilon < b$. That is, there is a single maximum provided that $\varepsilon < b$. For $\varepsilon > b$ the single maximum *bifurcates* into two maxima symmetrically located with respect to the r-axis. This happens, according to (20), for $r > \sqrt{b^2-\varepsilon^2}$.

3. The Dynamics of Evolution Strategies for FNIMs

Due to the complicated functional structure of $\mathrm{E}[F_4|\mathbf{x}]$, Eq. (14), and $\mathrm{Var}[F_4|\mathbf{x}]$, Eq. (18), one cannot apply the sphere model theory[6] in a simple fashion. Actually, $\mathrm{E}[F_4|\mathbf{x}]$ depends on two (aggregated) state variables, therefore, the dynamics and an underlying theory must contain at least two degrees of freedom. However, there are some clues that, qualitatively, the behaviour should share some common properties with the sphere model. At least the steady-state behaviour should exhibit some kind of residual localisation error for the optimiser: Because of Eq. (18) $\mathrm{Var}[F_4|\mathbf{x}] > 0$ does always hold (provided that $\varepsilon > 0$), even for the case $(r, x_N) = (0,0)$.

In order to get a certain feeling how the ES evolves on function F_4, ES runs have been performed. They are displayed in Figs. 9 – 11. The value of the parental vector $\mathbf{x}$ was randomly initialised on a hyper-sphere with radius $R^{(0)}$. Considering the shape of F_4 for vanishing noise (cf. Fig. 8, $\varepsilon = 0$), it becomes clear that under such random conditions the quadratic x_N part in (14) dominates resulting in large negative F_4-values. The ES increases these F_4-values very fast as can be seen in Fig. 9 (the average $\langle F_4 \rangle$ of the μ parent fitnesses is displayed). The fast F_4 increase stops when

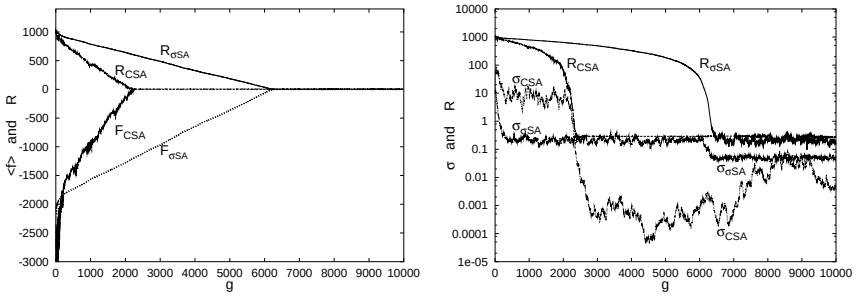


Fig. 9. The R- and σ-dynamics of a $(30/30_I, 60)$-ES on F_4 with $N = 20$, $a = 5$, $b = 0.5$, and $\varepsilon = 0.5$ using CSA and σSA, respectively. In the left figure, a linear vertical scale has been used while a logarithmic one in the right figure. The initial values are $R^{(0)} = 1000$ ($\mathbf{x}$-vector randomly initialised with $\|\mathbf{x}\| = R^{(0)}$) and $\sigma^{(0)} = 10$.

the "ridge"-like region has been reached. Then the dynamics changes into a linear one, the parental distance to the optimum $R^{(g)} = \|\mathbf{x}^{(g)}\|$ decreases obeying an almost perfect linear time law. The CSA-ES evolves faster to the steady-state than the σSA-ES. The steady-state is again characterised by a non-vanishing localisation error.

Figure 10, left-hand side, shows the approach to the steady-state considering the evolution of the x_N coordinate. Apart from the burst between

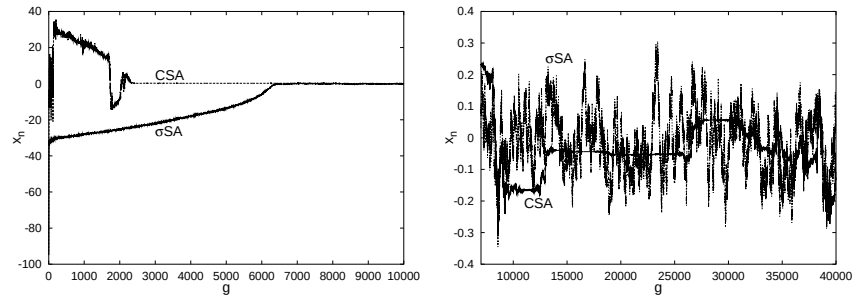


Fig. 10. Dynamics of the x_N-coordinate of the $(30/30_I, 60)$-ES run from Fig. 9 on F_4 using CSA and σSA, respectively. Left figure: transient phase; right figure: steady-state phase.

generation $g = 1700$ to 2200, there is nothing special with that coordinate. In the steady-state (right-hand side) it fluctuates around the zero line. However, consider Fig. 11, the difference to the ES run considered in Figs. 9 and 10 is the increased noise strength $\varepsilon = 0.75$. For this noise parameter, the

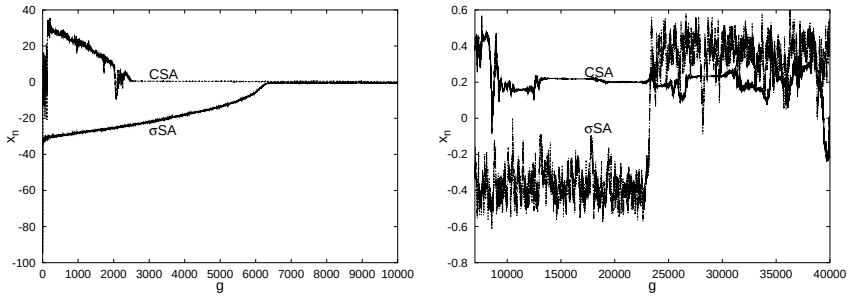


Fig. 11. The same conditions as in Fig. 10, but $\varepsilon = 0.75$ has been chosen.

σSA-ES exhibits a (random) periodic behaviour jumping back and forth between two attractors. Using a slightly smaller noise strength $\varepsilon = 0.7$, the time period gets smaller, Conversely, using larger noise strengths, e.g. $\varepsilon = 1.0$, the σSA-ES stays in one of the attracting regions forever (see the right-hand side in Fig. 12). While the σSA-ES exhibits periodic behaviour

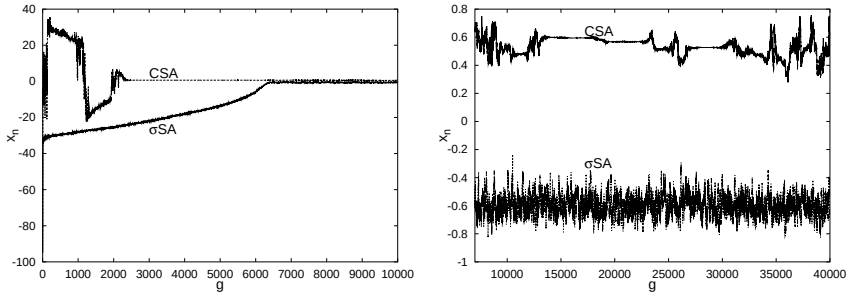


Fig. 12. The same conditions as in Fig. 10, but $\varepsilon = 1.0$ has been chosen.

for a certain noise level interval, the behaviour of the CSA-ES is more conservative. The reason for that lies in the small mutations strengths σ the CSA-ES randomly evolves when reaching the (almost selection neutral) steady-state. That is, unlike the σSA-ES, there is not enough mutation strength to push the ES system from one attractor to the other.

4. Steady State Behaviour of the Evolution Strategy on Function F_4

In order to estimate a lower bound on the residual localisation error we apply findings from the "standard" fitness noise model for evolution strategies, see the work by Arnold and Beyer[3,4] and Beyer *et al.*[6]. In the standard model the noise term in the fitness function is additive and normally distributed $\delta \sim N(0, \sigma_\delta^2)$. In order to apply the stability condition to ensure local convergence in the mean, which is given by (see Arnold and Beyer[4])

$$\sigma_\delta < \frac{R^2\,|\beta|}{N} \mu\, c_{\mu/\mu,\lambda}, \tag{21}$$

($R = R^{(g)}$ denotes the parental distance to the optimum, $c_{\mu/\mu,\lambda}$ the progress coefficient and β the factor of the sphere function) we first have to introduce appropriate sphere model approximations. Therefore, we have to neglect the influence of x_N in the denominator of Eq. (13). This step yields an ellipsoidal model. In a next *ad hoc* step, we assume that the eccentricity of the ellipsoid can be neglected. This leads to the desired sphere approximation (dropping the constant term a)

$$Q_{\mathrm{sp}}(\mathbf{x}) = -\|\mathbf{x}\|^2/b = -R^2/b. \tag{22}$$

In the variance expression (18) we neglect x_{N-1} and x_N totally, as a result

$$\sigma_\delta = \sqrt{\mathrm{Var}[f_2]} = \sqrt{2}\varepsilon^2/b. \tag{23}$$

Now, the evolution criterion (21) together with (23) and $\beta = -1/b$ is applied

$$\frac{\sqrt{2}\varepsilon^2}{b}\frac{Nb}{2R^2} < 2\mu c_{\mu/\mu,\lambda} \tag{24}$$

and finally solving for R, one obtains

$$R > R_\infty = \frac{\varepsilon}{2}\sqrt{\frac{\sqrt{2}N}{\mu c_{\mu/\mu,\lambda}}}. \tag{25}$$

Figure 13 shows the predictive quality of this formula. Even though the predictions seem to be relatively good, one should keep in mind that this result was obtained for a "moderate" b-value. One can easily violate the sphere condition by choosing more extreme b-values. Furthermore, considering larger ε-values, it appears that the asymptotic behaviour of (25) seems not to be correct.

Figure 13 (bottom) shows the behaviour of the mean value of the x_N coordinate in the steady-state. It reflects the behaviour observed in Figs. 10 – 12: Up to a certain ε the mean value is zero. After specific ε, the absolute mean values grow monotonously with the noise strength.

5. Extending F_4 to a More General Function Class

Considering broader and different classes of FNIM, respectively, is useful for at least two reasons. First, it broadens the view concerning the behaviour of EA in noisy settings, thus providing deeper insight in certain aspects of robust optimisation. Second, finding a special FNIM class which is especially suited for an analytical investigation of the ES behaviour on this functions. We discuss the properties of Function F_5 introduced in Section 2, Eq. (15) in this section in more detail. Using a slightly different notation, function F_5 can be re-written as:

$$F_5(\mathbf{x}) := a - \frac{r_1^2 + \sum_{i=N_1}^{N_2-1}(x_i + z_i)^2}{b + r_3^2} - r_3^2, \qquad b > 0, \tag{26}$$

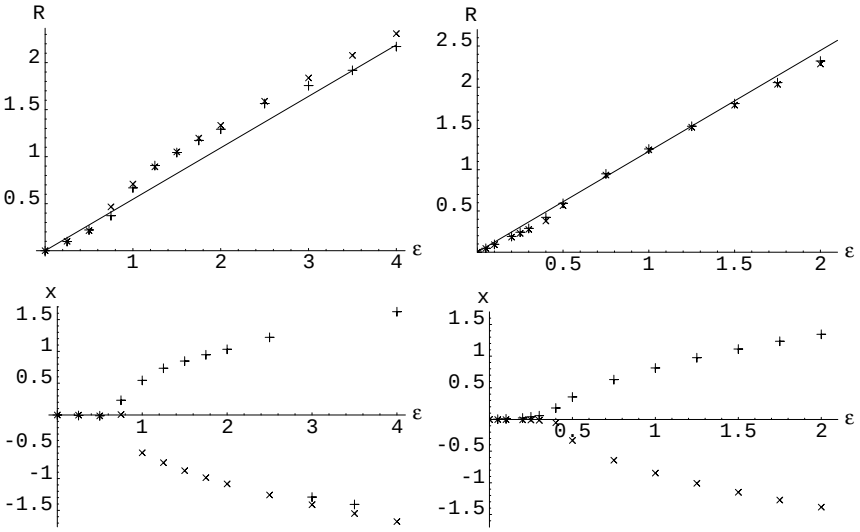


Fig. 13. Dependence of the mean value of the steady-state R (top figures) and of the steady-state x_N (bottom figures) on the noise strength ε. For the simulations a $(30/30_I, 60)$-ES has been used. Parameters of f_2 are $a = 5, b = 0.5$. The left figure was obtained for dimensionality $N = 20$ ($\varepsilon = 0$, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 3.5, 4) and the right figure for $N = 100$ ($\varepsilon = 0$, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2). The data points from the CSA-ES, displayed by "+"-symbols, are the average over generations $g = 3000$ (7000 for $N = 100$) to 40000 those of the σSA-ES, displayed by "×", are the average over generations $g = 8000$ (10000 for $N = 100$) to 40000. The linear curves represent Eq. (25).

where

$$r_1 := \sum_{i=1}^{N_1-1} x_i^2, \qquad r_3 := \sum_{i=N_2}^{N} x_i^2, \qquad 1 \le N_1 < N_2 \le N, \tag{27}$$

with

$$z_i \sim \mathcal{N}(0, \varepsilon^2). \tag{28}$$

We will discuss F_5 with respect to the mean value robustness. Since for each parameter space component with $i = N_1, \ldots, N_2 - 1$ the result of Eq. (14) holds similarly, we obtain for the expected value

$$\mathrm{E}[F_5|\mathbf{x}] = a - \frac{r_2^2 + (N_2 - N_1)\epsilon^2}{b + r_3^2} - r_3^2 \tag{29}$$

with r_2 defined by

$$r_2 := \sum_{i=1}^{N_2-1} x_i^2. \tag{30}$$

Comparing this expression with (17) we see the great similarity of both expressions. This also transfers to the optimal object parameter settings: $\mathrm{E}[F_5|\mathbf{x}]$ is locally optimised for $r_2 = 0$, that is, the first $N_2 - 1$ x_i coordinates must be zero. Applying the stationarity condition $\partial \mathrm{E}[F_5|\mathbf{x}]/\partial r_3 = 0$ for the coordinate aggregation r_3 yields

$$\tilde{r}_3 = 0, \qquad \text{for} \quad r_2 \le \sqrt{b^2 - (N_2 - N_1)\varepsilon^2}, \tag{31}$$

$$\tilde{r}_3 = \sqrt{\sqrt{r_2^2 + (N_2 - N_1)\varepsilon^2} - b}, \quad \text{for} \quad r_2 > \sqrt{b^2 - (N_2 - N_1)\varepsilon^2}. \tag{32}$$

Thus, we see that the global optimiser (use $r_2 = 0$ in (31)) depends on ε according to

$$\hat{\mathbf{x}} = \mathbf{0}, \qquad \text{for} \quad \varepsilon \le b/\sqrt{N_2 - N_1}, \tag{33}$$

$$\hat{\mathbf{x}} = (0, \ldots, 0, \hat{x}_{N_2}, \ldots, \hat{x}_N)^{\mathrm{T}}, \qquad \text{for} \quad \varepsilon > b/\sqrt{N_2 - N_1}, \tag{34}$$

where the $\hat{x}_{N_2}, \ldots, \hat{x}_N$ must conform the condition $\hat{r}_3 = \sqrt{\sqrt{N_2 - N_1}\varepsilon - b}$

$$\sum_{i=N_2}^{N} \hat{x}_i^2 = \sqrt{N_2 - N_1}\, \varepsilon - b. \tag{35}$$

This is an interesting result: For $N_2 < N$ there are no two global optimal solutions, but a $(N - N_2)$-dimensional manifold of solutions located on a (hyper-sphere) of radius $\sqrt{\sqrt{N_2 - N_1}\, \varepsilon - b}$ and origin at $(0, \ldots, 0)$ in the $\mathbb{R}^{N-N_2+1}$ subspace of the coordinates $x_{N_2}, \ldots, x_N$. This may be regarded

as an extreme form of multi-modality or as some kind of indifference. One can easily calculate the maximal expected fitness by inserting (33) and (35) into (29) with the result

$$\hat{\overline{F}}_5 = a - (N_2 - N_1)\,\varepsilon^2/b, \qquad \text{for} \quad \varepsilon \leq b/\sqrt{N_2 - N_1}, \tag{36}$$

$$\hat{\overline{F}}_5 = a + b - 2\sqrt{N_2 - N_1}\,\varepsilon, \quad \text{for} \quad \varepsilon > b/\sqrt{N_2 - N_1}. \tag{37}$$

Summing up, test function (26) allows for two types of noise-induced multi-modality:

(1) bimodality for $N_2 = N$ and
(2) infinite multi-modality for $N_2 < N$.

While the existence of the first case has already been confirmed by real ES runs in the last section, the second case is presented here. Figures 14 and 15 shows steady state values of ES runs on F_5 with $a = 5$, $b = 1$, $N = 40$, $N_1 = 23$, $N_2 = 39$. That is, r_3 is the aggregation of x_{N-1} and x_N. As one

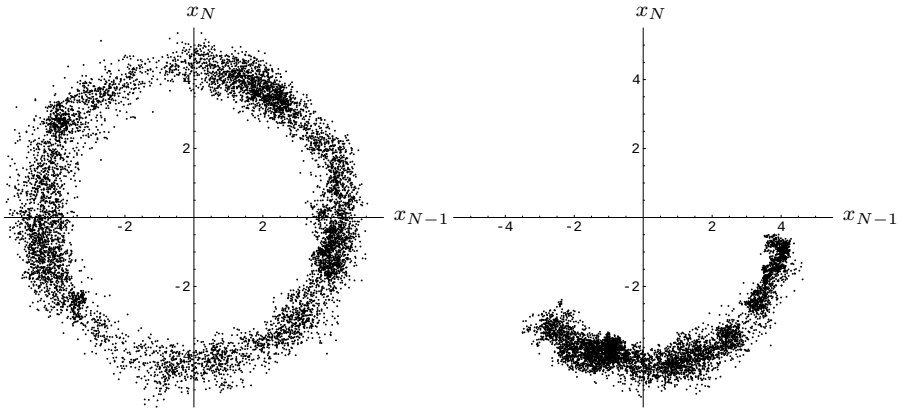


Fig. 14. On the distribution of the optimiser states of function F_5 ($a = 5$, $b = 1$, $N_1 = 23$, $N_2 = 39$, $N = 40$) in the vicinity of the steady state (8000 data points used) for $\varepsilon = 3$. The left figure was obtained using the $(5/5_I, 10)$-σSA-ES, the right figure using the $(5/5_I, 10)$-CSA-ES.

can see, if ε is sufficiently large, the behaviour predicted by (34) is observed. For small ε, x_{N-1} and x_N as well as the other x coordinates fluctuate around the zero state. This is in accordance with (33). The fluctuation around the optimiser state is the typical behaviour evolutionary algorithms do exhibit when evolving in a noisy environment. Note, while (35) describes

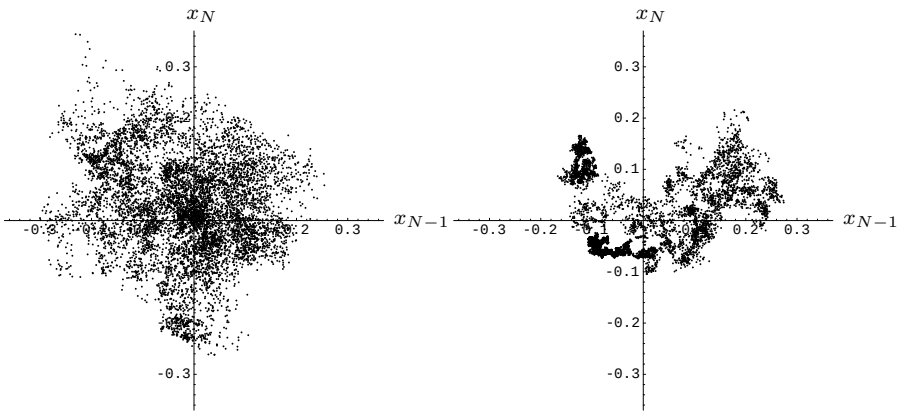


Fig. 15. On the distribution of the optimiser states of function F_5 ($a = 5$, $b = 1$, $N_1 = 23$, $N_2 = 39$, $N = 40$) in the vicinity of the steady state (8000 data points used) for $\varepsilon = 0.1$. The left figure was obtained using the $(5/5_I, 10)$-σSA-ES, the right figure using the $(5/5_I, 10)$-CSA-ES.

the optimum distance to the origin the x coordinates should realize, the actually observed mean value of the steady state r_3 deviates from this optimum. In the experiments conducted for Figs. 14 and 15 one measures $r_3 \approx 4.3$ for the $(5/5_I, 10)$-σSA-ES and $r_3 \approx 4.1$ for the CSA-ES ($\varepsilon = 3$). The optimum value, however, is $\hat{r}_3 = \sqrt{11} \approx 3.32$. The actually observed mean value is a result of the evolutionary algorithms and depends on the strategy parameters. Its calculation is still a pending problem.

It is interesting to notice the different behaviours the σSA-ES and the CSA-ES exhibit for the case $\hat{r}_3 > 0$. As one can see in Fig. 14, the CSA-ES (right figure) does not occupy the whole circle (notice, this is a plot of a single ES run). Depending on the initial values chosen, it is likely to observe this typical pattern. The reason for this observation can be traced back to the evolution of the mutation strength in CSA-ES under the influence of heavy noise also observed in the evolution dynamics of F_4 and other test functions (see Sendhoff *et al.*[12]). Under heavy noise, the CSA-ES first reduces the mutation strength σ and then it more or less performs a random walk in the σ values, but keeping these values small. However, small σ values result in small changes of the object parameters. (This can also be observed in the right picture of Fig. 15.) In other words, the CSA-ES becomes less explorative. Under such conditions, the CSA-ES is not able to explore the whole $r_3 = const.$ subspace efficiently. This is in contrast to the σSA-ES which does not reduce the mutation strength in such situations. Which one

of the two behaviours is more desirable, however, is application dependent. Therefore, we cannot give a definitive answer as to the question which of the two σ-adaptation rules should be preferred.

6. Summary and Conclusion

In this chapter, we discussed the effect of noise on the search space in evolutionary algorithms. We introduced three main characteristics which go beyond the simple additive noise model. The *expectation-variance trade-off*, the *topology changing* functions and the *functions with noise induced multi-modality (FNIM)*. The last two function classes have the remarkable property of qualitatively changing the topology under the influence of noise. For the FNIM's the change from uni-modal to bi-modal or multi-modal fitness landscapes which we termed a bifurcation process using the analogue from nonlinear dynamics, occurs when a measure for the robustness or stability of the solution is used for the fitness.

We derived the conditions for bifurcation and empirically analysed the influence of the topological change of the fitness landscape on the behaviour of two types of evolution strategies, the cumulative step-size adaptation method and the "standard", mutative self-adaptation method. Whereas the later one exhibits periodic behaviour for a certain noise level interval, the CSA method tends to converge to one of the two optima. Although the proposed class of test functions is rather different from the sphere model, we were able to transfer some results from sphere model analysis at least qualitatively. In order to extend this analysis to a more quantitative one, which could help to give some insight into appropriate ranges of parameters like population size and selection pressure, requires a substantial step in the theory of evolutionary algorithms. However, at the same time, it can serve as an interesting test problem in this domain, because of its "natural" transition from uni-modal to bi-modal characteristics.

The extension of FNIMs $F_{3,4}$ to a more general class of functions F_5 in the last section demonstrated an additional important aspect for comparing CSA-ES and σSA-ES. The fact that for function F_5 the bifurcation leads to a manifold of optimal solutions (instead of to two isolated optima) highlighted the different behaviour of the two types of evolution strategies for exploring selectively neutral parts in search space. However, as we have seen already for function F_2 the fact that optima are not necessarily unique is by no means restricted to noisy optimisation tasks. Whether we should demand from an optimiser to identify one single solution as fast as possible

(CSA-ES) or the identification of the whole set of optimal solutions (σSA-ES) has to be answered for each specific application separately. At the same time, looking at the area of multi-objective optimisation where the notion of a space of optimal solutions, i.e. the Pareto space, occurs naturally, the identification of all solutions might be desirable.

Acknowledgements

B. Sendhoff and M. Olhofer thank E. Körner for his support. H.-G. Beyer acknowledges support from the Collaborative Research Center SFB 531 sponsored by the Deutsche Forschungsgemeinschaft (DFG).

Appendix A. Description of the Evolution Strategies

The σ self-adaptation technique is based on the coupled inheritance of object and strategy parameters. Using the notation

$$\langle \mathbf{a} \rangle^{(g)} := \frac{1}{\mu} \sum_{m=1}^{\mu} \mathbf{a}_{m;\lambda}^{(g)} \tag{A.1}$$

for intermediate recombination (centroid calculation, i.e., averaging over the $\mathbf{a}$ parameters of the μ best offspring individuals), the $(\mu/\mu_I, \lambda)$-σSA-ES can be expressed in "offspring notation"

$$\forall l = 1, \ldots, \lambda : \begin{cases} \sigma_l^{(g+1)} := \langle \sigma \rangle^{(g)} \mathrm{e}^{\tau \mathcal{N}_l(0,1)} \\ \mathbf{y}_l^{(g+1)} := \langle \mathbf{y} \rangle^{(g)} + \sigma_l^{(g+1)} \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{1}). \end{cases} \tag{A.2}$$

As learning parameter $\tau = 1/\sqrt{N}$ has been chosen in the simulations.

While in evolutionary self-adaptive ES each individual get its own set of endogenous strategy parameters, cumulative step-size adaptation uses a single mutation strength parameter σ per generation to produce all the offspring. This σ is updated by a deterministic rule which is controlled by certain statistics gathered over the course of generations. The statistics used is the so-called (normalised) cumulative path-length $\mathbf{s}$. If $\|\mathbf{s}\|$ is greater than the expected length of a random path, σ is increased. In the opposite situation, σ is decreased. The update rule reads

$$\begin{aligned} &\forall l = 1, \ldots, \lambda : \mathbf{y}_l^{(g+1)} := \langle \mathbf{y} \rangle^{(g)} + \sigma^{(g)} \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{1}) \\ &\mathbf{s}^{(g+1)} := (1-c)\mathbf{s}^{(g)} + \sqrt{(2-c)c} \frac{\sqrt{\mu}}{\sigma^{(g)}} \left(\langle \mathbf{y} \rangle^{(g+1)} - \langle \mathbf{y} \rangle^{(g)} \right) \\ &\sigma^{(g+1)} := \sigma^{(g)} \exp\left(\frac{\|\mathbf{s}^{(g+1)}\| - \overline{\chi}_N}{D \overline{\chi}_N} \right) \end{aligned} \tag{A.3}$$

where $\mathbf{s}^{(0)} = \mathbf{0}$ is chosen initally. The recommended standard settings for the cumulation parameter c and the damping constant D are used, i.e., $c = 1/\sqrt{N}$ and $D = \sqrt{N}$, see also Hansen and Ostermeier[10].

References

1. A.N. Aizawa and B.W. Wah. Scheduling of genetic algorithms in a noisy environment. *Evolutionary Computation*, 2(2):97–122, 1994.
2. D. Arnold. *Noisy Optimization with Evolution Strategies*. Kluwer Academic Publishers, 2002.
3. D. V. Arnold and H.-G. Beyer. Local performace of the $(\mu/\mu_I, \lambda)$-ES in a noisy environment. In W. Martin and W. Spears, editors, *Foundations of Genetic Algorithms, 6*, pages 127–141. Morgan Kaufmann, 2001.
4. D. V. Arnold and H.-G. Beyer. Performance analysis of evolution strategies with multi-recombination in high-dimensional $I\!R^N$-search spaces disturbed by noise. *Theoretical Computer Science*, 289:629–647, 2002.
5. H.-G. Beyer. Toward a theory of evolution strategies: Some asymptotical results from the $(1,^+ \lambda)$-theory. *Evolutionary Computation*, 1(2):165–188, 1993.
6. H.-G. Beyer, M. Olhofer, and B. Sendhoff. On the behavior of $(\mu/\mu_I, \lambda)$-ES optimizing functions disturbed by generalized noise. In K. A. De Jong, R. Poli, and J. E. Rowe, editors, *Foundations of Genetic Algorithms VII*, pages 307–328, 2002.
7. J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research. Springer Verlag, 1997.
8. J. Branke. *Evolutionary Optimization in Dynamic Environments*. Kluwer Academic Publishers, 2001.
9. J.M. Fitzpatrick and J.J. Grefenstette. Genetic algorithms in noisy environments. In P. Langley, editor, *Machine Learning: Special Issue on Genetic Algorithms*, volume 3, pages 101–120. Kluwer Academic Publishers, 1988.
10. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–196, 2001.
11. Y. Jin and B. Sendhoff. Trade-off between optimality and robustness: An evolutionary multiobjective approach. In C. M. Fonseca, P. J. Fleming, E. Zitzler, K. Deb, and L. Thiele, editors, *Evolutionary Multi-Criterion Optimization*, pages 237–251. Springer Verlag, 2003.
12. B. Sendhoff, H.-G. Beyer, and M. Olhofer. On noise induced multi-modality in evolutionary algorithms. In L. Wang, K.C. Tan, T. Furuhashi, J.-H. Kim, and X. Yao, editors, *Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution And Learning - SEAL*, volume 1, pages 219–224, 2002.
13. S. Tsutsui and A. Gosh. Genetic algorithms with a robust solution searching scheme. *IEEE Trans. on Evolutionary Computation*, 1(3):201–208, 1997.
14. D. Wiesmann, U. Hammel, and T. Bäck. Robust design of multilayer optical coatings by means of evolutionary algorithms. *IEEE Trans. on Evolutionary Computation*, 2(4):162–167, 1998.