PREPRINT

A Time Travel to the Early Theory of Evolution Strategies

The 15th March 1965 was the deadline for Hans-Paul Schwefel to submit the final version of his diploma thesis to the Hermann Föttinger institute for fluid mechanics at the Technical University of Berlin. He was requested by his advisor Prof. Dr.-Ing. Rudolf Wille to work on the answers to three questions:

1. **Is the evolution strategy superior to classical optimization strategies in case of many variables?**
2. **How does the evolution strategy behave in case of noise?**
3. **Is there a theoretical justification for using stochastically large, medium and small mutation steps? And if so, what is their optimal distribution?**

There are well-founded reasons why the results are hardly known to the scientific community that is engaged in the field of Evolutionary Computation: Diploma theses are practically not available via public libraries, the diploma thesis is written in German, … and there was no WWW in 1965.

This article is supposed to put Hans-Paul Schwefel's findings in a nutshell in order to initiate a dissemination of these early contributions to the theory of evolution strategies.

Starting Position

The starting position is the assumption that one has to cope with experimental optimization, i.e., there is no mathematical description or a simulator for the system to be optimized: you are optimizing a real object at hardware level! Since the interrelationships between the variable input parameters and the dependent output behavior are unknown, we encounter a black box situation in the cybernetic sense: In case of a closed system with high complexity the only thing you can do is the measuring of the input/output relations.
In experimental optimization the measure for assessing the behavior or quality of the real object varies but the main optimization loop remains the same: You change some settings (i.e. input parameters) of the system according to some strategy and then you measure some physical quantity that reflects the quality of the system. Depending on the measuring the underlying optimization strategy dictates which settings are to be altered and how.

Analysis of Gradient Strategy

The model for analyzing different optimization strategies is developed on pp. 6-12. Let x be a real vector of dimension $n \geq 2$ that represents the input parameters. These parameters should be changed according to some desired (unknown) direction in parameter space. If f(x) measures the quality to be maximized then the best direction in a reasonably sized vicinity of the current position x is the direction given by the (unknown) gradient $\nabla f(x)$. Let $\delta > 0$ be the distance made into the desired direction. Then $\varphi = \delta / m$ describes the velocity toward the optimum parameter setting, where m denotes the number of measurements required in the course of a single iteration of the optimization strategy.
The gradient strategy is certainly a classical optimization method. If it is supposed to be used in experimental optimization it has to be modified: Since gradients are not available in our

black box scenario they must be approximated via $\partial f / \partial x_i \approx \Delta f / \Delta x_i$. Thus, at least $m = n + 1$ measurements are required for an approximation of the gradient. Since the approximation is only valid in a vicinity of order Δx_i, the step size s must not exceed this value. As a consequence, the progress δ toward the desired direction is simply s and therefore we obtain $\varphi = \delta / m = s / (n + 1)$ for the velocity toward the optimum. If the measurements are inaccurate due to noise, additional measurements must be made for averaging out the errors. In this case the velocity decreases to $\varphi = s / (nk + k)$ where k denotes the number of measurements per setting.

Analysis of Cybernetic Evolution Strategy

After the analysis of the gradient strategy the strategy of *cybernetic evolution* is introduced (pp. 13 – 18) next. According to Schwefel, the term *cybernetic evolution* is used to denote the transfer of principles of natural evolution to procedures for attaining a given goal, e.g. in scientific and technical research. Although this term is uncommon nowadays, its meaning has been transferred to the term *evolutionary computation* in the meantime.

Prior to the formal analysis of the evolution strategy (ES) some obvious differences to the gradient strategy (GS) are listed: The ES is not restricted to a single path through the parameter space and a comparison of two different states is based on an ordinal instead of a cardinal scale. This property makes the ES less sensitive to perturbations than the GS which uses the values of the measurements to determine the next step to be taken.

A Markov Model with Discrete Mutations

Let $X^{(k)}$ be the parameter setting (i.e., the individual) of dimension n at iteration $k \geq 0$. The new candidate solution $Y^{(k)}$ is obtained via $Y^{(k)} = X^{(k)} + s^{(k)} U^{(k)}$, where $s^{(k)} > 0$ is the scalar step size and $\{U^{(k)}\}_{k \geq 0}$ is a sequence of independent and identically distributed random vectors of size n. The components U_i of U are independent with distribution (p. 19)

$$P(U_i = -1) = P(U_i = 0) = P(U_i = 1) = 1/3.$$

Let $s^{(k)} = s > 0$ be constant. As a consequence, the ES operates on a lattice (embedded in $\Re^n$) whose granularity is determined by the size of s. Evidently, the early ES used in practice or analyzed theoretically was optimizing over *discrete* structures – this widely unknown fact is in diametrical opposition to the common myth that ES are only used for optimization over continuous sets!

Assume that the step size s is small enough such that the objective function $f: \Re^n \rightarrow \Re$ (i.e., the quality measure) can be approximated linearly in the sense of a Taylor expansion at the current position $X^{(k)} \in \Re^n$. In two-dimensional space (n = 2) the objective function is locally approximated by $f(x) = x_1 \cos \alpha + x_2 \sin \alpha$, where angle α specifies the desired direction toward better solutions. As can be seen in fig. 1, for every angle α there is a transition graph whose vertices (the squares) denote the positions on the lattice and whose directed edges represent transitions to vertices with better objective function value.

If the ES is in state i then there are 3^n possible new positions caused by mutation. The transition probability is $m_{ij} = 1/3^n$ for those states j and $m_{ij} = 0$ otherwise. The selection operation compares the objective function values of i and j. If state j is better than i (i.e., if there is an edge from i to j in the graph of fig. 1), then the probability of accepting state j is given by $s_{ij} = 1$ and $s_{ij} = 0$ otherwise. If there are noisy measurements of the objective function value then $s_{ij} \in [0, 1] \subset \Re$. In any case, the transition probability from some state i to state j is $p_{ij} = m_{ij} \cdot s_{ij}$ for $i \neq j$ and $p_{ii} = 1 - \sum_j p_{ij}$ with $j \neq i$. Most likely, this is the first time-

homogeneous Markov chain model (pp. 20-25) of an evolutionary algorithm! The term *Markov model*, however, was never used in the diploma thesis.

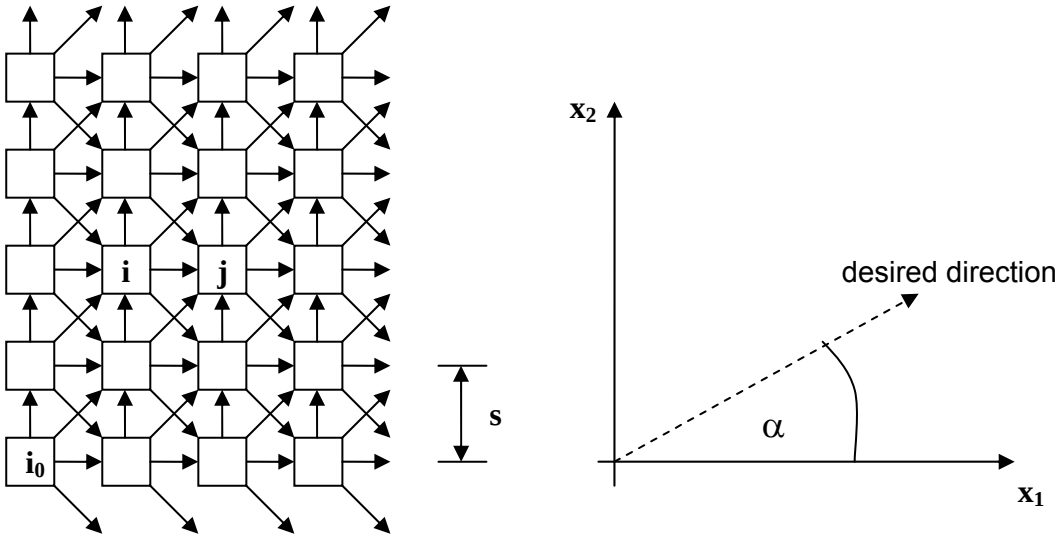

Fig. 1: *An embedded transition graph (left) where directed edges represent potential transitions to better states. The existence of the edges and their directions depend on the angle α that determines the direction to better solutions (right).* Reproduced from [1], p. 21.

Owing to this Markov model the progress was determined numerically on a Zuse Z23 computer. Not surprisingly, the practicability of this approach was limited since the number of states and therefore the size of the matrices became intractably large – and this remains true for contemporary computers!

As a consequence, a different analytical approach was developed (pp. 26-30). The numerically tedious calculation of the occupancy frequencies in the Markov model can be circumvented by calculating the expected progress directly: Let the ES with discrete mutations optimize the (unbounded) linear function $f(x) = c'x$ for some vector c.

If $n = 2$ and $0 < \alpha < 45°$ the expected progress is $\varphi(\alpha) = s\,(3 \cos\alpha + \sin\alpha) / 9$ for some $s > 0$ and a specific angle α. Integration and averaging over all angles $\alpha \in (0, 45°)$ leads to the general expected progress $\varphi = 4s\,(1+\sqrt{2}) / (9\pi) \approx 0.342$ s. Compared to $\varphi = s / 3 \approx 0.333$ s for the gradient strategy it is easily seen that ES beats GS for $n = 2$. Using a lower bound on φ for $n > 2$ (credited to Ingo Rechenberg) it is shown numerically (p. 30) that the ES is the better than the GS the larger is the dimension n.

Expected Progress along a Parabolic Ridge

This test problem was used to analyze the behavior of the ES in a nonlinear environment (pp. 35-46). Again, the ES used discrete mutation steps of fixed size. Using the same approach as before (i.e., calculate the expected progress *directly*) Schwefel shows that GS and ES behave similar if the step size s is in the range being valid for gradient approximations. In contrast to the GS the ES is not fooled if the step sizes are made much larger. In this case the ES is up to one order of magnitude quicker than the GS. But there is a limit on the size of s at which the ES is blocked: This kind of stagnation distant from a local optimum occurs sooner or later depending on s. A simulation on a Zuse Z23 computer with 1.3×10^6 mutations in 100 hours computing time confirmed the necessity of a formal analysis of this phenomenon.

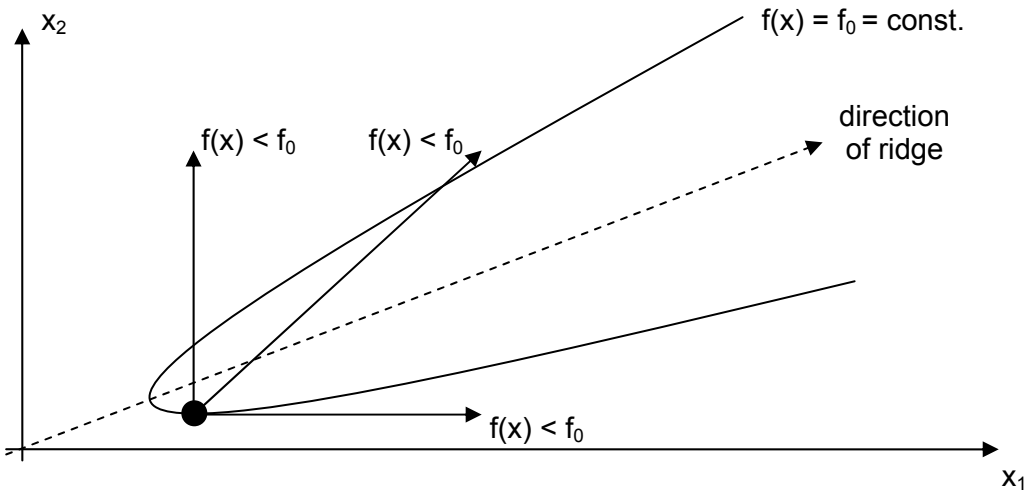

Fig. 2: *Blocking caused by constant step sizes: The ES cannot generate better solutions from its current position (black spot). Every solution within the parabola is better than the current position for this maximization problem.* Reproduced from [1], p. 41.

It is proven that blocking is caused by *discrete* mutation steps (also see fig. 2). Schwefel succeeded in deriving analytical expressions for the regions of blocking depending on s for n = 2. It turned out that the region of blocking becomes smaller the finer is the lattice (i.e., the smaller is s). This leads to the conclusion that these regions of blocking will vanish if the step size can be made arbitrarily small. Thus, one needs some kind of step size control!

Expected Progress for the Rotational Parabola

The test problem termed *rotational parabola* is nowadays known as the *sphere model*. The expected progress was calculated (pp. 47-50) for the ES with discrete mutations with fixed step size for n = 2 and n = 3. The analysis revealed that the progress decreases when approaching the vicinity of the optimum and that the ES got stuck prematurely prior to reaching the optimum. The invariability of the step sizes was identified as the reason for this unpleasant behavior.

Continuous Step Size Distributions

The analysis of the progress rates for the parabolic ridge and the rotational parabola had revealed that the ES is handicapped by its method to generate new candidate solutions: If the step size is constant and discrete the ES can stagnate somewhere in the search space without reaching a local optimum. There can only be one conclusion: The ES must be endowed with a variable and continuous step size distribution!

The Gaussian law of errors was the background for choosing the normal distribution for mutations. It was shown (pp. 51-53) how the expected step size can be controlled by the standard deviation σ by deriving the probability density of the stochastic step size (that obeys a $\chi_n(\sigma)$-distribution).

A formal analysis is still possible with this kind of step sizes (pp. 54-59). The expected progress is determined for the unbounded linear case with and without noise in two-dimensional space. Finally, this is the version of the ES as we know it today with the exception that the explicit control of σ was not addressed here: It took some years until a satisfactory solution was found; the development of effective and efficient step size controls is still an active area of research in the EC community.

Conclusions

(1) The ES beats the GS especially in high dimensions. (2) The ES works even in the presence of noise. (3) Constant discrete step sizes lead to premature stagnation of the search. (4) Step sizes should be variable, continuously distributed, and controllable in future versions of ES.

Outlook

Future versions of ES should include further principles from natural evolution beyond mutation and selection. Actually, the principle of isolation is emphasized and a sketch of evolutionary algorithms with periodically communicating subpopulations is given. This model became known as the migration model in the EC community decades later when parallel computers became affordable.

In 1965 this and other ideas were just dreams. Many of them came true. Others still await their fulfilment …

Bibliography

[1] Hans-Paul Schwefel: Cybernetic Evolution as Strategy for Experimental Research in Fluid Mechanics (in German). Diploma Thesis, Hermann-Föttinger Institute for Fluid Mechanics, Technical University Berlin, March 1965.