Natural Evolution and Collective Optimum-Seeking

Hans-Paul Schwefel

Department of Computer Science, University of Dortmund, P.O. Box 50 05 00, D-4600 Dortmund 50, Germany

Abstract

On the one hand many people admire the often strikingly efficient results of organic evolution, on the other hand, however, they presuppose mutation and selection to be a rather prodigal and unefficient trial-and-error strategy like Monte-Carlo sampling. Taking into account the parallel processing of a heterogeneous population and sexual propagation with recombination as well as the endogenous adaptation of strategy characteristics, simulated evolution reveals a couple of interesting, sometimes surprising, properties of nature's learning-by-doing algorithm. 'Survival of the fittest', often taken as Darwin's view, turns out to be a bad advice. Forgetting, i.e. individual death, and even regression show up to be necessary ingredients of the life game. Whether the process should be termed gradualistic or punctualistic, is a matter of the observer's point of view. He even might observe 'long waves'.

1. INTRODUCTION

Evolution can be looked at from a large variety of positions. Beginning with the closest physico-analytic viewpoint, one might focus attention to the molecular and cellular processes. A more distant point of view centers on the behaviour of populations or species. Another difference emerges from whether one emphasizes the homeostatic aspect of the adaptation to a given environment, which is more relevant in the short term, or the euphemistic view of development to the more complex, higher, or even better in the long term.

The instruments used here will be a macroscope and a time accelerator. Moreover, for methodological reasons, an optimistic point of view will be shared by comparing macro-evolution with iterative optimization, or, more adequately, with permanent meliorization techniques, i.e. hill-climbing or ridge-following procedures. By means of a simple algorithmic formulation of the main evolutionary principles, it is possible to reveal some properties of the process which in some cases are striking at the first glance. These findings may not only be helpful for better understanding 'nature's intelligence' but also be beneficial for global long-term planning and other groping-in-the-dark situations.

2. MODELLING EVOLUTION

Ashby's homeostat [1] was a device which should find back to a feasible state by a sequence of random trials, uniformly distributed over a given parameter space. Many

people have made the mistake of thinking of mutations as 'pure' random trials. A couple of them was malignant. They wanted to show that evolution theory never will be able to explain how 'nature' found a way to complex living beings within about 10^{17} seconds - the age of our globe. Montroll's [7] random walk paradigm, on the other hand, neglects the selection principle of evolution. Both mutation and selection (chance and necessity) are the first principles, which, of course, have to be programmed properly.

Broadly accepted hereditary evidence has led to the proverb 'The apple does not fall far off from the tree'. A better model of mutations therefore is a normal distribution for phenotypic changements between generations, its maximum being centered at the respective ancestor's position. The rôle of chance in such a model is not explicative, however, but only descriptive. An important question now is the suitable size of the standard deviation(s) of the changes, which may be addressed as mean step size(s) from one generation to the next. This question arises with all optimization or meliorization schemes.

Modelling the selection principle is far more easy, as it seems at first. 'Survival of the fittest' is the maxim which some people derived from Darwin's observations. Some evolution programmers have taken it for granted: According to a given selection criterion, a descendant is rejected if its vitality is less than that of its ancestor, the ancestor otherwise. This scheme may be called a $(1 + 1)$ or two membered evolution strategy (ES in the following), resembling the 'struggle for life' between one ancestor *and* one descendant. Rechenberg [10] has derived theoretical results for the convergence velocity of that process in an n-dimensional parameter space. Most important was his finding that for an endless ridge following situation as well as for a minimum (or maximum) approaching situation the convergence rate is inversely proportional to the number of parameters. Distances growing with the square root of n, the number of iterations or generations needed to proceed from one to the other arbitrary point in space, increases with $O(\sqrt{n})$ only, and not geometrically as in the case of simple Monte-Carlo strategies.

This type of creeping random search strategy (see e.g. Brooks [5], Schumer and Steiglitz [11], or Rastrigin [9]) was first devised for experimental optimization, where measurement inaccuracies drop out one-variable-at-a-time and gradient-following procedures due to their inability of non-local operation.

Bremermann's 'simulated evolution' [4] does not differ so much from Rechenberg's as e. g. G.E.P. Box's 'Evolutionary Operation' EVOP [2] does, an experimental design technique, and the so-called Simplex and Complex strategies of Nelder and Mead [8] and M.J. Box [3] for numerical optimization. Whereas random trials were vividly rejected by G.E.P. Box, he centered several experiments (principally at the same time) in a deterministic way around the position of the current best point in a low-dimensional parameter space. The best of all then is taken as the center of the next trial series. Nelder and Mead, and M.J. Box, however, using a polyhedron for placing the trials, reject the worst position and find a new one by reflecting the worst with respect to the center of the remaining points of the simplex or complex.

The first concept may be called a $(1 + \lambda)$, the latter a $(\mu + 1)$ evolutionary scheme, μ denoting the number of parents, λ the number of children within one generation. More general, therefore, is a $(\mu + \lambda)$ scheme with μ ancestors which have λ descendants, the μ best *of all* become parents of the next generation. The fact that individual life times

are limited is reflected by the (μ, λ) version, first introduced by Schwefel [12]. Now the μ parents are no longer included into the selection, thus λ must be greater than μ. Theoretical results so long are available for the $(1+\lambda)$ and $(1, \lambda)$ evolution strategies only. All further observations in the following, therefore, were found by computer simulation only.

3. SELF-ADAPTATION OF STRATEGY PARAMETERS

As for all optimization techniques, the appropriate step size adjustment is of crucial importance. Rechenberg found that there is a 'window' of one decade only within which the $(1+1)$ evolution strategy has a reasonable convergence velocity. He devised a simple rule for exogenously adjusting a near optimum performance of the process, i.e. to observe and control the success probability which should be kept in the vicinity of one success among five trials. This advice is good for many but not all situations. Moreover, it does not give any hints to adapt the standard deviations of the parameter changements individually. Some may be too large, others too small, at the same time. Only within the multimembered strategy, one can include the step size or even different step sizes (mutation rates) into the set of the individual's genes and adapt them endogenously. There is some evidence that by means of repair enzymes the effective mutation rates are controlled, the rate of premutations due to environmental conditions being constant over long periods.

Let us think at first, however, of one common step size for all object parameters. Within a $(1, \lambda)$ ES the correct step size turns out to be even more important than within a $(1 + \lambda)$ version [12]. In the first case regression takes place instead of progress when the step size is too large, whereas stagnation is the worst case in the latter. At a first glance, therefore, 'survival' of an ancestor might be a good advice. Simulation results, however, show that the opposite is true. This is the first surprise. Figure 1 demonstrates the difference between $a\ (1 + 10)$ and $a\ (1, 10)$ ES when minimizing the function

$$F_1 = \sum_{i=1}^{n} x_i^2 \quad \text{with } n = 30 \text{ real parameters } x_i. \tag{1}$$

The 'progress' was measured in terms of $log(\sqrt{F^0/F^g})$, where F^0 denotes the start value, F^g the current value of the objective function at generation/iteration g.

The number of variables, n, was taken to be as large as 30 in order to avoid improper conclusions. In lower dimensional parameter spaces nearly every strategy may achieve good results. One common step size σ (or standard deviation, more precisely) for all x_i is changed by mutation, i.e. by multiplying the ancestor's value with a random number, drawn from a logarithmic normal distribution in order to avoid exogenous drift. The $(1, 10)$ strategy turns out to be superior. An explanation for this surprising fact is the following: If an ancestor happens to arrive at a superior position, this might be - by chance - in spite of a non-optimum step size, or a step size which is not suitable for further generations. The $(1 + \lambda)$ scheme preserves the unsuitable step size as long as with

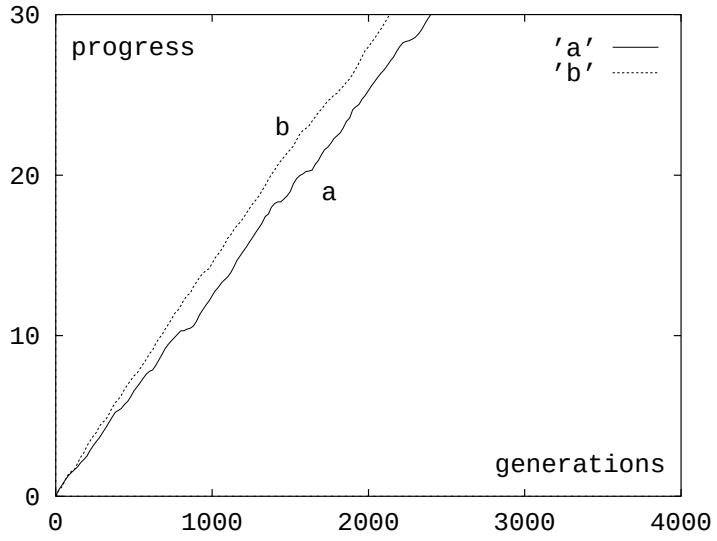

Figure 1: Self-learning of one common mutation step size σ for function F_1, $n = 30$; a) (1+10) evolution strategy; b) (1,10) ES.

it a further success is placed. This leads to periods of stagnation. Within a $(1, \lambda)$ ES the good position, occasionally won with an unsuitable step size, is lost, together with the latter, during the next generation. This short term regression, however, enhances the long term velocity of the whole process by a stronger selection with respect to the suitable step size (strategy parameter). In other words: Forgetting is as important as learning, the former must be seen as a necessary integral part of the latter. One might interpret the fact of an inherent finite life time of living beings (preprogrammed maximum number of cell divisions) as an appropriate measure of nature to overcome the difficulties of undeserved success - or, in a changing environment, of forgetting outdated 'knowledge'.

4. COLLECTIVE LEARNING OF PROPER SCALINGS

In most cases it is not sufficient to adapt one common step size for all object parameters. For an objective function like

$$F_2 = \sum_{i=1}^{n} (i\ x_i^2) \quad \text{with } n = 30 \text{ again,} \tag{2}$$

for example, individual standard deviations σ_i corresponding to the x_i and appropriately scaled, could speed up the progress rate considerably. To achieve this kind of flexibility within the multimembered evolution strategy, each individual is characterized by a set of n step sizes in addition to the n object parameters. They are mutated by multiplication with two random factors, one being common for all step sizes as before, the other acting

individually, however. Thus general and specific scalings can be learned at the same time. Operating with an $(1, \lambda)$ strategy - however large λ may be - leads to a second surprise: this kind of process does not work at all, it gets stuck prematurely by approaching a relative optimum in a lower dimensional space. The reason is rather simple: As said above, the convergence rate is inversely proportional to n, the dimension of the parameter space. Descendants operating in a subspace by sharply reducing some of the step sizes have a short-term advantage. Selecting the fittest descendant to become the one and only parent of the next generation, is counterproductive in the long-term, as was the possibility of survival of the ancestor.

Figure 2 demonstrates how to overcome the difficulty. If more than one, i.e. not only the best of the descendants, become parents of the next generation and recombination by sexual propagation takes place, i.e. mixing of the information gathered by different individuals during the course of evolution, then over-adaptation and consecutive stagnation can be overcome. Now the convergence rate steeply goes up with the population size. On a conventional one-processor computer the parallelism of that scheme cannot be realized properly, but multi-processor machines are entering the market place now. That is why all following figures show the progress over the number of generations and not over computing time. The total number of individuals must not increase with the number of object parameters, however, if the problem complexity remains constant like with objective functions F_1 and F_2 from above.

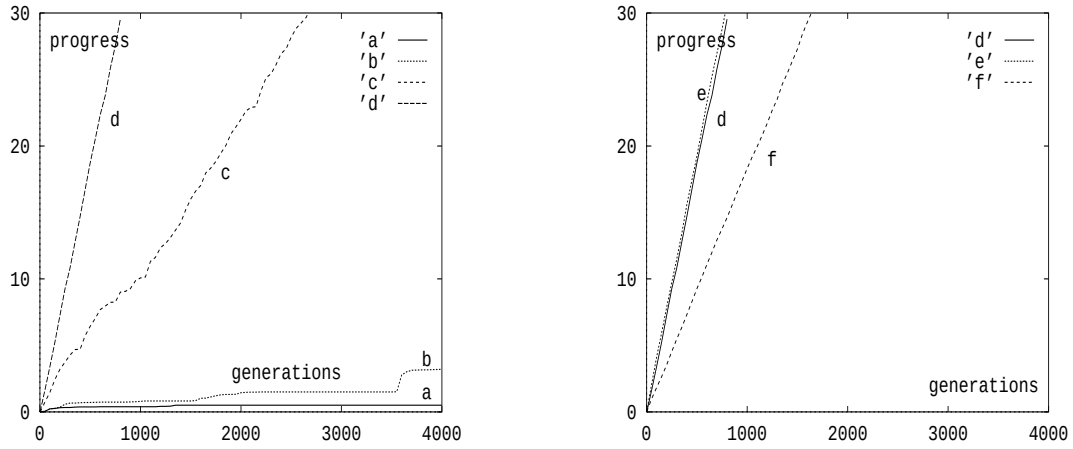

Figure 2: Learning of the scaling; a) (1,10) ES; b) (3,10) ES with recombination; c) (6,30) ES with recombination; d) (15,100) ES with recombination; e) (1,100) ES and f) (15,100) ES both without recombination but with prefixed optimum scalings.

The overwhelming success of recombination demonstrated here, may explain the early appearance of sexual propagation on earth. But it is unprobable that only the additional variability provokes that success. A better explanation might be the following: The typical situation during the meliorization process is following a ridge. Within a population some individuals have a position on one side, others on the other side of the ridge. Mixing

genetic information is a means of riding the ridge more efficiently. A similar argument holds for mixing the step size knowlege: Individuals on one side of the ridge have 'internal models' (made up of the set of step size relations) of the response surface which are different from those on the other side. Even if both models are wrong for the long term, since both may be locally adapted only (if the 'model' learned is not a law of nature), some 'mean' model (or better: hypothesis) may turn out to be more useful for the future. One may interpret that phenomenon by saying that 'natural intelligence' is distributed.

Now the question of the appropriate selection pressure prevails: How many of the descendants should be selected as new parents. The answer is given by figure 3. All other conditions being held constant, including the number of descendants $\lambda = 100$ within one generation, only μ, the number of parents or 'survivors' from λ new-born genotypes, was changed. Three cases were investigated for function F_2.

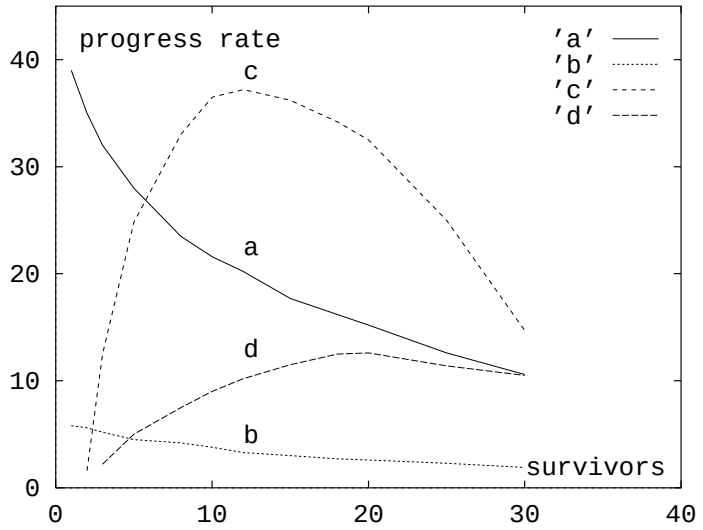

Figure 3: Comparing progress rates per 1000 generations of different $(\mu, 100)$ ESs over μ, the number of parents (survivors of $\lambda = 100$ descendants); a) with prefixed optimum scaling ($\sigma_i^0 = c/i^{\frac{1}{2}}$); b) with prefixed arbitrary scaling ($\sigma_i^0 = c$); c) with adaptive scaling by means of individual learning plus recombination; d) with adaptive scaling by recombination only.

Whereas in both cases a) and b) $\mu = 1$ is the best choice, in a learning situation (case c and d), it is better, even necessary, to increase μ far beyond 1. The diagram moreover demonstrates the effectivity of the collective learning process. Under proper conditions nearly the same convergence rate as with total knowledge of the optimum scaling can be achieved - even and only with lower selection pressure. This is the third surprise. Recombination alone (case d) is not as successful, however, as together with individual mutations (case c) of the genetically transmitted information about the different step sizes (which represent the internal model of the current/local environment).

5. LEARNING OF A METRIC BY MEANS OF THE EPIGENETIC APPARATUS

Topologies of vitality response surfaces normally are not as simple as assumed above. The next possible complication is to incline the main axes of the hyper-ellipsoids which form the subspaces $F = \text{const}$. Now scaling alone does not help in achieving optimum performance. What can nature do, what has it done, to overcome the difficulty? In many cases one has found that a single gene influences several phenotypic characteristics (pleiotropy) and vice versa (polygeny). These are the two sides of the same coin, which is correlation, the perhaps best known example of it being allometric growth. The transmission mechanism between genotype and phenotype, called epigenetic apparatus, in a first order may be approximated by linear correlation. In addition to individual step sizes, now correlation coefficients or angles of inclination of the ellipsoid, forming the surface of constant probability density of a mutation, had to be learned.

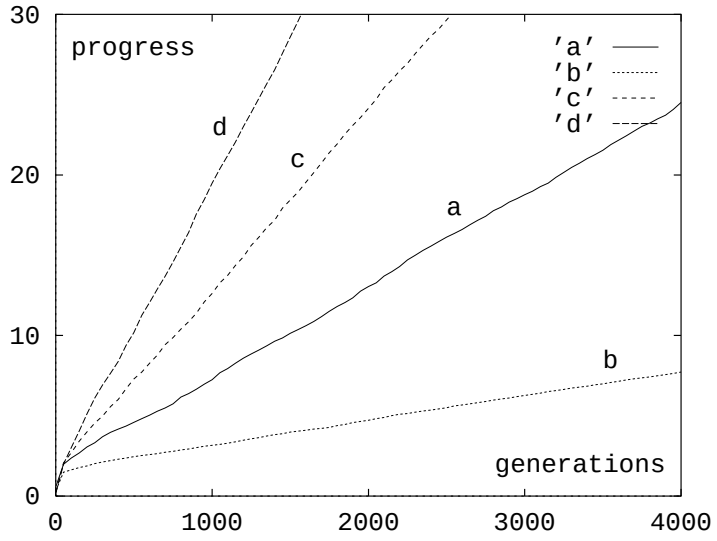

Figure 4: Learning of a metric; a) (1,100) ES with constant but arbitrary scaling; b) (15,100) ES under the same conditions, i.e. without recombination; c) (15,100) ES with recombination and adaptive individual step sizes; d) (15,100) ES as before, with additional learning of linear correlations between the phenotypic mutations.

Figure 4 shows first results for the objective function

$$F_3 = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j^2 \right) \quad \text{with } n = 10. \tag{3}$$

Four cases were simulated, three of them corresponding to those of figure 2. In both cases c) and d), recombination, intermediate for both the object parameters and the step sizes, was used. It is obvious that these sampling conditions bear a variety of possibilities with respect to the mutabilities of step sizes and correlation angles so that simulations c) and d) might not yet respresent the best choices. Nevertheless the results show how much may be gained in terms of progress velocity by allowing to learn a simple 'internal model' of the topology of the environment, the 'real world'. Nevertheless, such a model in most cases will not be a correct 'theory', but simply a useful local or temporal hypothesis. Especially and again, no single individual has the best long-term knowledge about the 'correct' world model. This knowledge is partial, temporal only, and distributed.

6. GRADUALISM OR PUNCTUALISM, AND SO-CALLED LONG WAVES

Up until now all figures showing the evolutionary progress over time or generations were drawn for objective function values only every 50th generation and for the mean of the (parents) population moreover. If we have a closer look by picking out one of the decision variables, e.g. x_{15} for function F_2, and look at it at every generation (figure 5, curve a) then the picture reveals more details. It depends on the density of the historical

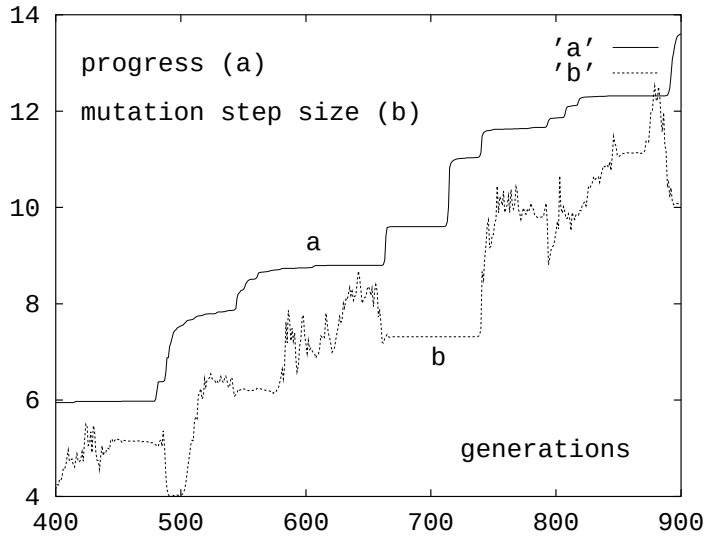

Figure 5: Time cut of one of the variables within a (3,100)- ES for F_2.

record whether we may speak of a gradual or a punctualistic process (see e.g. Stanley [13]). Due to the fact that the objective function (curve a) depends on many parameters, a single one of them, like x_{15} (curve b), must not resemble the progress as a whole. Great success in one direction forgives regression in others. And, for sure, looking at some arbitrary cut of the time record, one might get the impression of stochastically disturbed 'long waves' (normally three) with a more or less constant period. Even more aggregated subobjectives, like the GNP of a nation, could exhibit such behaviour, if the underlying process operates left of the maximum of curve c) in diagram 3, as was the case in figure 4 with $\mu = 3$ and $\lambda = 100$.

7. CONCLUSIONS

Many people today, when speaking about long term planning, environmental forecasting, technology assessment etc., are embarrassed by the degree of our ignorance. Very often then they speak of the uncertainties involved. But looking more closely, isn't it a matter of fact that there are, at the same time and with access to the same data, different certainties, i.e. different interpretations of the past and different aspirations for the future, or, in other words, different 'internal models' of the world? In the light of the simulation results above, one should appreciate, not regret, that. Due to the findings of a rather new field of science, i.e. nonlinear dynamics, we must admit that knowledge about the long-term future is principally unavailable for an open dissipative system operating far off from equilibria. We really are groping in the dark [6]. Therefore we should not try to establish one best model of the world, but make the best of the different individual ones of the ridge we are trying to follow without clearvoyance. Even if all of them were wrong they might be worthwhile to be recombined with each other. Instead of relying upon too strong competition, which leads to stagnation, as we have seen, we better should agree upon further experiments in cooperation and have pity with the loosers by means of solidarity, simply because they, too, enhance our knowledge by exploring our environment.

REFERENCES

1 W.R. Ashby, Design for a Brain, 2nd ed., Wiley, New York, 1960.

2 G.E.P. Box and N.R. Draper, Evolutionary Operation: A statistical method for process improvement, Wiley, New York, 1969.

3 M.J. Box, A New Method of Constrained Optimization and a Comparison with Other Methods, Computer Journal, 8 (1965) 42-52.

4 H.J. Bremermann, Numerical Optimization Procedures Derived from Biological Evolution Processes. In: Cybernetic Problems in Bionics. (H.L. Oestreicher and D.R. Moore, eds.), Gordon and Breach, New York, 1968.

5 S.H. Brooks, A Discussion of Random Methods for Seeking Maxima, Oper. Res., 6 (1958) 244-251.

6 D. Meadows, J. Richardson, and G. Bruckmann, Groping in the Dark - The First Decade of Global Modelling, Wiley, Chichester, 1982.

7 E.W. Montroll and K.E. Shuler, Dynamics of Technological Evolution: Random Walk Model for the Research Enterprise, Proc. Natl. Acad. Sci. USA 76, (1979) 6030-6034.

8 J.A. Nelder and R. Mead, A Simplex Method for Function Minimization, Computer Journal, 7 (1965) 308-313.

9 L.A. Rastrigin, Sluchainyi Poisk v Zadachakh Optimisatsii Mnogoparametricheskikh Sistem, Zinatne, Riga, 1965.

10 I. Rechenberg, Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution, Frommann-Holzboog, Stuttgart, 1973.

11 M.A. Schumer and K. Steiglitz, Adaptive Step Size Random Search, IEEE Trans., AC-13 (1968), 270-276.

12 H.-P. Schwefel, Numerical Optimization of Computer Models, Wiley, Chichester, 1981.

13 S.M. Stanley, The New Evolutionary Timetable, Basic Books, New York, 1981.