A study of the effects of dimensionality on stochastic hill climbers and estimation of distribution algorithms

Laurent Grosset[1,2], Rodolphe Le Riche[2], and Raphael T. Haftka[1]

[1] Mechanical and Aerospace Engineering Department, University of Florida, USA
[2] CNRS URA 1884 / SMS, École des Mines de Saint Étienne, France

Abstract. One of the most important features of an optimization method is its response to an increase in the number of variables, n. Random stochastic hill climber (SHC) and univariate marginal distribution algorithms (UMDA) are two fundamentally different stochastic optimizers. SHC proceeds with local perturbations while UMDA infers and uses a global probability density. The response to dimensionality of the two methods is compared both numerically and theoretically on unimodal functions. SHC response is $\mathcal{O}(n \ln n)$, while UMDA response ranges from $\mathcal{O}(\sqrt(n) \ln(n))$ to $\mathcal{O}(n \ln(n))$. On two test problems whose sizes go up to 7^{200}, SHC is faster than UMDA.

1 Introduction

Random stochastic hill climber (SHC) and univariate marginal distribution algorithms (UMDA, Mühlenbein and Paaß, 1996) are two fundamentally different stochastic optimizers. SHC proceeds with local perturbations while UMDA infers and uses a global probability density. It is important to understand how these two different search processes scale with the number of variables n.

Previous contributions have analyzed stochastic hill climbers and population-based evolutionary algorithms on specific objective functions. Garnier et al. (1999) computed the expected first hitting time and its variance for two variants of stochastic hill-climbers. Droste et al. (2002) extended the estimation of the running time of $\mathcal{O}(n \ln n)$ for a (1+1)-evolution strategy (ES) to general linear functions. Several studies have investigated the benefits of a population. For example, SHC and genetic algorithms have been compared in Mitchell et al. (1994) on the Royal Road function. Jansen and Wegener (2001) presented a family of functions for which it can be proven that populations accelerate convergence to the optimum even without crossover. He and Yao (2002) used a Markov chain analysis to estimate the time complexity of evolutionary algorithms (EA) for various problems and showed that a population is beneficial for some multimodal problems. Comparisons between single point and population-based evolution strategies on Long-Path problems are given in Garnier and Kallel (2000). There has been little work on the time complexity of estimation of distribution algorithms. The convergence of estimation of distribution algorithms has been studied in Mühlenbein et al. (1999). Experimental comparisons were conducted by Pelikan et al. (2000). However, there is a lack of theoretical results on the relative performances of SHC and UMDA.

The present paper is a numerical and analytical investigation of the effect of the number of non-binary variables on the performance of SHC and UMDA for two separable, unimodal functions. The topology of the functions is purposely simple so that the effect of dimensionality can be isolated. The performance of the methods for a multimodal, separable function is also discussed.

2 Presentation of the algorithms

We consider the problem of maximizing some fitness function F over a design space $D = A^n$, where A is an alphabet of cardinality c. SHC searches the space by choosing an initial point $x = (x_1, x_2, \ldots, x_n)$ at random and applying random perturbations to it: the algorithm changes the value of one variable chosen at random to an adjacent value and accepts the new point only if it improves the fitness function. Perturbations are applied until no more progress is observed.

UMDA is a simple form of estimation of distribution algorithms (EDA). EDAs use populations of m points to infer the distribution $p(x)$ of good points. By a succession of sampling and selection steps, the distribution converges toward regions of increasing fitness evaluation, eventually yielding the optimum. In UMDA, distributions are expressed as products of univariate marginal distributions, leading to the following update rule for the distributions:

$$p(x, t+1) = \prod_{i=1}^{n} p^s(x_i, t) \tag{1}$$

where $p(x, t+1)$ refers to the search distribution at time $t+1$ and $p^s(x_i, t)$ designates the univariate distribution of the variable x_i in selected points at time t. In this work, truncation selection of ratio τ was used.

The algorithm can be summarized as follows:

1. set t = 0,
2. initialize the search distribution $p(x, 0)$,
3. create m points by sampling from $p(x, t)$,
4. select the τm best individuals based on the fitness function,
5. estimate the univariate marginal distributions $p^s(x_i, t)$, and calculate $p(x, t+1)$ from Equation (1),
6. go to 3.

3 Numerical experiments

Three test problems are considered. The first two problems, “Max A_{11}” and “Vibration”, are separable and unimodal functions. The third, “Min A_{66}”, is a multimodal and separable function. All three problems have a physical meaning in terms of composite laminate design.

A composite laminate is made of layers of fiber reinforced material (plies), and its response is determined by many factors including the number of plies, the fiber volume fraction, the fiber and matrix properties, etc. In this work, we consider only laminates that have a fixed number of plies n, as shown in Figure 1. The goal of laminate optimization is to determine the angles x_1 to x_n that maximize some objective function. Even though fiber angles can take any value between 0° and 90°, the set of acceptable values is typically limited to a small number of discrete values for manufacturing requirements. In the problems considered in this paper, the angles can take seven values from 0° to 90° in 15° steps. Note that the angles values are ordered, i.e. they are not symbolic, hence the notion of distance underlying SHC is well defined.

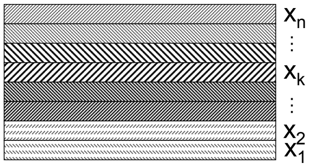

Fig. 1: Laminate cross-section

3.1 Max A_{11} problem

The first problem consisted in maximizing the longitudinal in-plane stiffness A_{11} of a symmetric balanced laminate[3], which is expressed by:

$$A_{11} = U_1\, h + 4U_2 \sum_{k=1}^{n} t_k \cos 2x_k + 4U_3 \sum_{k=1}^{n} t_k \cos 4x_k \tag{2}$$

where U_1, U_2 and U_3 are material constants, n is the total number of plies, t_k the thickness of the k^{th} ply, and h the total thickness of the laminate.

The fitness function of this problem is a sum of functions of one variable only, so that they do not interact with one another, and UMDA is expected to converge to the global optimum $x_i^* = 0°$, $i = 1, n$. The function A_{11} is unimodal, so that SHC will also yield x^*.

The algorithms were applied to the problem for five different numbers of variables $n = 12, 20, 50, 100, 200$. The selection ratio τ of the UMDA was kept constant at $\tau = 0.3$ (as recommended in Mühlenbein et al. (1999)). Several

[3] A laminate is symmetric if the ply orientations are symmetric about its mid-plane. It is balanced when for each $+\theta$ ply, there is a $-\theta$ in the laminate. Therefore, a balanced and symmetric laminate has n unknown ply orientations and $4n$ constitutive plies.

population sizes (50, 100, 500 and 1000) were tried in order to obtain an efficient scheme and allow a fair comparison with SHC.

Two criteria were used to compare the algorithms' performance: the number of analyses required to reach 80% reliability (defined as the probability of finding the optimum, estimated over 50 independent runs), and the number of analyses needed until the average best fitness reaches 98% of the optimal fitness. Clearly, these two criteria provide only a partial picture of the algorithms' relative performances, and other criteria may be used. However they measure two quantities that are critical in optimization algorithms: the maximum fitness criterion is an indication of the velocity of the convergence in the vicinity of the optimum, while the reliability criterion measures the algorithm effectiveness at finding the true optimum.

Figure 2 presents the number of evaluations to 98% of the maximum fitness against the number of variables for SHC and four different population sizes of UMDA. Clearly, SHC converges faster than UMDA for all the numbers of variables investigated. The evolution of the cost for SHC appears to be linear, which confirms the results reported in section 4 and in Pelikan et al. (2000) for the *Onemax* problem. For UMDA with a given population size, the number of evaluations needed to reach 98% of the optimal fitness increases sub-linearly. Larger populations are more expensive, but smaller population can fail to converge for large n. This is the case when a population of 50 individuals is used to solve the Max A_{11} problem and $n \geq 50$, where the average maximum fitness never reaches 98% of the maximum fitness. This can be explained by the fact that when smaller populations are used, the chance of losing particular values of the variables is higher, which prevents the algorithm form finding the optimum, as will be discussed in section 4.2.

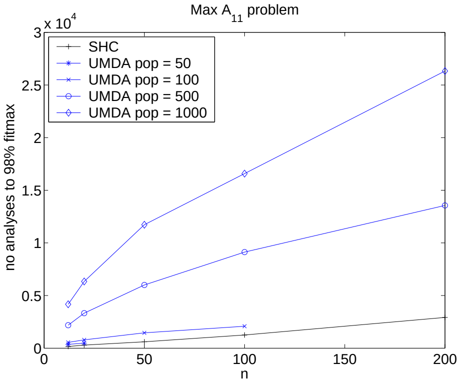

Fig. 2: Number of analyses until the average maximum fitness reaches 98% of the optimal fitness, Max A_{11} problem.

The effect of the loss of variable values for small populations is visible in the reliability: for each problem size n, there exists a minimum population size below which 80% reliability is never reached because of premature convergence of the distributions. This minimum population size was $m^* = 100$ for $n = 12$, $m^* = 500$ for $n = 20$, 50 and 100, $m^* = 1000$ for $n = 200$.

In the Max A_{11} problem, all the variables are interchangeable, as the order of the variables does not affect the fitness function. Consequently, the expected value of the univariate distribution of all the variables is the same. Figure 3 shows the evolution of the probability distributions of the variables corresponding to the innermost ply for the case $n = 100$, $m = 500$. Starting from a uniform distribution over the seven values, the optimum value ($p(x_1 = 0°, t) = 1$, $p(x_1 = i, t) = 0,\ i \neq 0°$) gradually takes over the whole distribution.

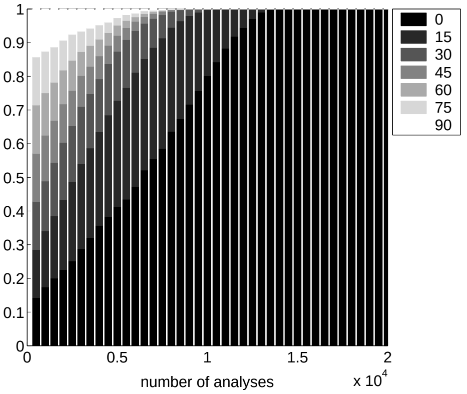

Fig. 3: Evolution of the probability distribution for the innermost ply.

3.2 Vibration problem

The second problem consisted in maximizing the first natural frequency of a simply supported rectangular laminated plate. This problem has independent variables, but exhibits a hierarchical structure in which the contribution of each variable to the fitness function depends on its position.

The first natural frequency of a simply supported rectangular plate is proportional to the square root of the following expression:

$$F = \frac{D_{11}}{L^4} + \frac{2(D_{12} + 2D_{66})}{L^2 W^2} + \frac{D_{22}}{W^4} \tag{3}$$

where L and W are the length and width of the plate.

The bending stiffness coefficients D_{ij} are obtained by:

$$D_{11} = U_1 \frac{h^3}{12} + \frac{4}{3} U_2 \sum_{k=1}^{n} t_k (3z_k^2 - t_k^2) \cos 2x_k + \frac{4}{3} U_3 \sum_{k=1}^{n} t_k (3z_k^2 - t_k^2) \cos 4x_k \tag{4}$$

$$D_{22} = U_1 \frac{h^3}{12} - \frac{4}{3} U_2 \sum_{k=1}^{n} t_k (3z_k^2 - t_k^2) \cos 2x_k + \frac{4}{3} U_3 \sum_{k=1}^{n} t_k (3z_k^2 - t_k^2) \cos 4x_k \quad (5)$$

$$D_{66} = U_5 \frac{h^3}{12} - \frac{4}{3} U_3 \sum_{k=1}^{n} t_k (3z_k^2 - t_k^2) \cos 4x_k \quad (6)$$

$$D_{12} = U_4 \frac{h^3}{12} - \frac{4}{3} U_3 \sum_{k=1}^{n} t_k (3z_k^2 - t_k^2) \cos 4x_k \quad (7)$$

where the U terms are material constants and z_k refers to the position of the k^{th} ply in the laminate.

The effect of z is to give a hierarchical structure to the problem because plies located in outer layer have more weight than those located in inner layers. As a result, the distributions corresponding to the outermost plies are expected to converge faster than those corresponding to inner plies. The optimum laminate for this problem was $x_i^* = 60°$, $i = 1, n$.

The two algorithms were applied to the vibration problem for the five problem sizes n used in the previous section. The number of evaluations necessary for the average maximum fitness function to reach 98% of the optimal fitness is shown in Figure 4. The results are similar to those obtained for the Max A_{11} problem. The cost of SHC is still linear in the number of variables. However, the number of evaluations needed by the two algorithms to reach 98% of the maximum fitness is smaller than on the Max A_{11} problem. For instance, for $n = 12$, SHC needs 64 evaluations on the vibration problem, against 160 on the Max A_{11} problem. In the case $n = 200$, it requires 1402 analyses for the vibration problem, against 2923 for the Max A_{11} problem. Similarly, the number of analyses needed by UMDA with a population of 1000 individuals decreases from 4175 analyses to 2028 analyses for $n = 12$ and from 26322 analyses to 17531 analyses for $n = 200$. The faster convergence toward high fitness regions for the second problem can be explained by the fact that a large part of the response is governed by outermost plies, so that most of the fitness improvement can be achieved by determining the value of these influential plies. In addition, the fact that the optimum angle (60°) is close to the center of the domain helps SHC by reducing the average number of steps it has to take.

If the hierarchical structure of the problem allows a rapid convergence to high fitness regions, it also causes numerical difficulties for UMDA. The convergence of the probability distribution for the innermost and outermost plies for the case $n = 100$, $m = 500$ are presented in Figures 5 and 6, respectively. On this hierarchical problem, it appears very clearly that the algorithm proceeds from the outside to the inside, starting with the more influential variables, and determining the inner variables only at the very end.

This mechanism is responsible for the loss of variable values for the less influential inner plies: in the early stages of the search, the selection of good individuals is mainly determined by the outermost variables. As a result, individuals which contain the optimum value of the outermost plies but not of the innermost plies get selected, potentially leading to the disappearance of these

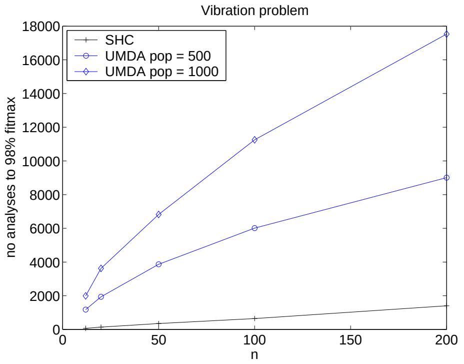

Fig. 4: Number of analyses until the average maximum fitness reaches 98% of the optimal fitness, vibration problem.

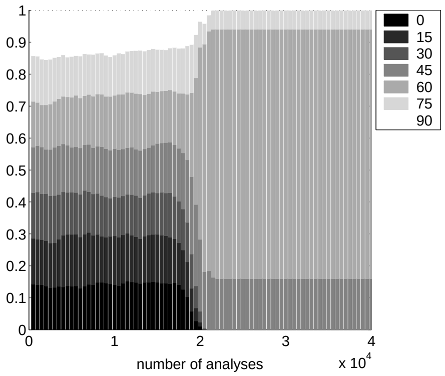

Fig. 5: Evolution of the probability distribution for the innermost ply.

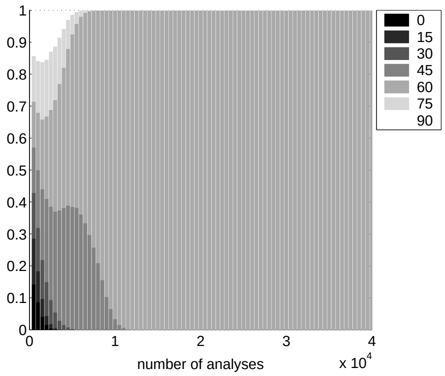

Fig. 6: Evolution of the probability distribution for the outermost ply.

values in the distribution if too small a population is used. In order to prevent the loss of values, we observed that larger populations have to be used. The minimum population sizes for this problem were $m^* = 500$ for $n = 12$ and $n = 20$, $m^* = 1000$ for $n = 50$. In the cases $n = 100$ and $n = 200$, the algorithm did not reach 80% reliability for the population sizes tested within the maximum number of 40,000 analyses used in this work.

SHC, however, is unaffected by the problem's hierarchical structure because it merely compares two neighboring points. Figure 7 shows the two performance measures on the two problems. Both the number of evaluations to 98% of the maximum fitness and the number of evaluations until 80% reliability is achieved are lower for the vibration problem than for the Max A_{11} problem.

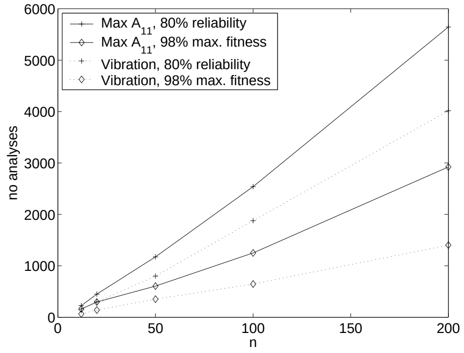

Fig. 7: Comparison of the performance of SHC on the Max A_{11} problem and the vibration problem.

3.3 A multimodal, separable function

SHC is misled by local minima, while the relationship between UMDA and the objective function is more complex. It is nevertheless known that, if the objective function is separable[4] and the population size is larger than m^*, UMDA converges to the global optima (Mühlenbein et al. (1999)). An example of such an objective function where the reliability of UMDA tends to 1 while that of SHC is nearly 0 is the minimization of the in-plane shear stiffness of a composite laminate, A_{66}, over ply orientations that are comprised between 0° and 75°,

$$\min_{0^\circ \ \leq \ x_i \ \leq \ 75^\circ} A_{66} \ , \tag{8}$$

where

$$A_{66} \ = \ U_5 h - U_3 \sum_{i=1}^{n} t_i \cos(4x_i) \ . \tag{9}$$

The global optimum is $x_i^* \ = \ 0^\circ$, $i \ = \ 1, n$. Replacing any of the x_i^* with 75° creates a local optimum whose basin of attraction starts at $x \ = \ 45^\circ$. For an n-dimensional case, there are 2^n local optima. Numerical experiments with $n \ = \ 12$ averaged over 50 runs confirm that the SHC reliability is 0 (it is theoretically $(45/75)^{12} \ = \ 2.10^{-3}$) while that of an UMDA with $m \ = \ 500$ reaches 1 after 3500 analyses. Figure 8 shows the evolution of the probability distribution of the outermost ply $p(x_n, t)$. It is interesting to note that in the early stages of the search, both the probability of 75° and the probability of 0° (the two local optima) increase. But after about 2000 analyses, the probability of 75° starts to decrease and the algorithm converges to the global optimum.

[4] A more general result is given in Mühlenbein et al. (1999) where the convergence of "Factorized Decomposition Algorithm" (FDA) to the optima is proved for additively decomposed functions. Separable functions are a special case of additively decomposed functions and UMDA is the corresponding simplified FDA.

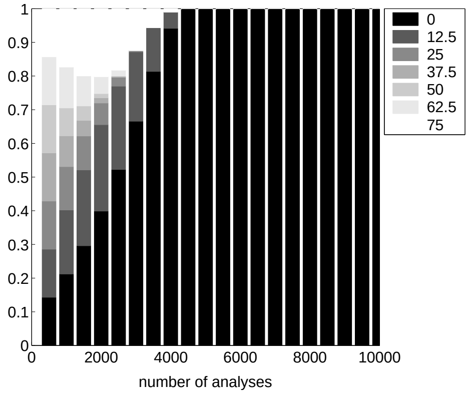

Fig. 8: Evolution of the probability distribution for the outermost ply, Min A_{66} problem.

4 Elements of theoretical explanations

The numerical experiments that were just described are now theoretically analyzed by calculating expected convergence times. There are n variables that can take c discrete values.

4.1 Convergence time of the SHC

The following analysis considers a stochastic hill climber (SHC) operating on a unimodal function. If the SHC is at a point where k out of the n variables are correctly set, the expected time before one of the non optimal variable is perturbed is $n/(n-k)$. The random perturbation can then take the variable closer to the optimum or not, with probabilities $1/2$ (neglecting distortions due to limits on the values). The expected time for one beneficial step is then $2n/(n-k)$. Let d_i denote the average distance between the i-th variables of a random point and the optimum,

$$d_i \;=\; \frac{1}{c}\sum_{j=1}^{c} |x_i^j - x_i^*| \;, \tag{10}$$

where x_i^j is the j-th possible value of the i-th variable. In the cases presented here, all variables values at the optimum are the same ($x_i^* \;=\; 0°$ for $\max A_{11}$ and $\min A_{66}$ or $x_i^* \;=\; 45°$ for the vibration problem), therefore the average distance to the optimum is the same for all variables i, $d_i \;=\; d$ (but it varies with the problem). At each variable that is not correctly set, an average of d steps in the right direction is needed to reach the optimum. By summing the expected times of each beneficial step, one obtains the expected time to locate the optimum from a random starting point that has k optimal variables

$$T_k \;=\; \sum_{i=k}^{n-1} \frac{2dn}{n-i} \;. \tag{11}$$

T_k can now be averaged over all random starting points, which yields the expected convergence time of an SHC on a unimodal function,

$$T_{SHC} \;=\; \frac{1}{2^n}\sum_{k=0}^{n} C_n^k T_k \;=\; \frac{nd}{2^{n-1}}\sum_{k=0}^{n} C_n^k \sum_{i=k}^{n-1} \frac{1}{n-i}\ . \tag{12}$$

It can be shown Garnier et al. (1999) that Equation (12) yields a convergence time of order $\mathcal{O}(n \ln n)$ for large n. Estimated and measured convergence times are compared for both the max A_{11} and the vibration problems in Figure 9. The dependency of the number of function evaluations in terms of the dimension n is correctly predicted. Because d is smaller in the vibration problem than in the max A_{11} problem, convergence is faster in the first case, which is also correctly predicted.

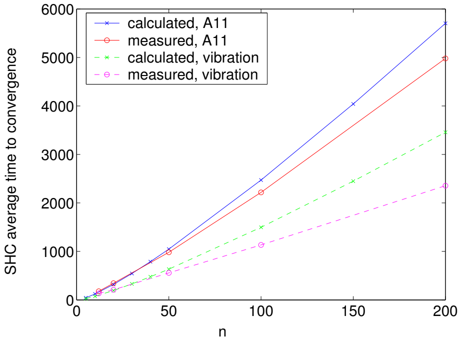

Fig. 9: Estimated and measured convergence times of SHC for the max A_{11} and the vibration problems.

4.2 Convergence time of the UMDA

A univariate marginal distribution algorithm is now considered. In Mühlenbein et al. (1999), the behavior of an UMDA with truncation selection is studied on two model functions, the *Onemax* and the *Int*, that have common features with the max A_{11} and the vibration problems. In the *Onemax* problem, the number of 1's in a binary string is maximized. Like in max A_{11}, the function is separable, and each variable has the same contribution to the objective function. The *Int* function,

$$Int \;=\; \sum_{i=1}^{n} 2^{i-1} x_i\ , \tag{13}$$

is also maximized on binary strings. Like in the vibration problem, the function is separable and there is a gradual influence of the variables on the function. In the vibration problem however, the difference in sensitivity of the objective function to each variable is lower than in Int. If the population size, m, is larger than a critical value m^*, it is shown in Mühlenbein et al. (1999) that the expected number of generations to convergence[5], N_g, is

$$N_g \approx \mathcal{O}(\sqrt{n}) \qquad \text{for } Onemax \tag{14}$$
$$N_g \approx \mathcal{O}(n) \qquad \text{for } Int\ . \tag{15}$$

The larger number of generations seen on Int is due to the different weights variables have on the objective function. For a truncation rate $\tau = 0.5$, the first selection is exclusively based on x_n, the second selection is based on x_{n-1}, ... The discovery of the optimum is sequential in variable values, whereas some level of parallelism can be achieved on less hierarchical objective functions.

The expected number of objective function evaluations to convergence is

$$N_f = N_g m^* \ . \tag{16}$$

No analytical expression for m^* was given in Mühlenbein et al. (1999). An approximation to m^*, $\widehat{m^*}$ is now proposed based on the initial random population sampling, and neglecting variable values lost during selection. The probability that a given variable value is not represented in the population is $((c-1)/c)^m$. The probability that the values making up the optimum, x^*, have at least a sample in the initial population is

$$P_{pop} = \left(1 - \left(\frac{c-1}{c}\right)^m\right)^n \ . \tag{17}$$

For a given P_{pop} (typically close to 1), the critical population size is estimated from Equation (17),

$$\widehat{m^*} = \frac{\ln(n/(1-P_{pop}))}{\ln(c/(c-1))} \approx \mathcal{O}(\ln(n)) \ . \tag{18}$$

From Equations (14) to (18), the order of magnitude of the number of evaluations to convergence is

$$N_f \approx \mathcal{O}(\sqrt{n}\ln(n)) \qquad \text{for } Onemax \text{ and} \tag{19}$$
$$N_f \approx \mathcal{O}(n\ln(n)) \qquad \text{for } Int\ . \tag{20}$$

These orders of magnitude agree well with the UMDA convergence trends seen on the max A_{11} and vibration tests in section 3. Note however that, because the vibration problem is not as hierarchical as Int in terms of variables, its experimental convergence time behaves more like $\sqrt{n}\ln(n)$ than $n\ln(n)$.

[5] Following Mühlenbein et al. (1999), convergence time is defined here as the time when $p(x^*, N_g) = 1$.

5 Concluding remarks

The asymptotic time to convergence of the SHC on a unimodal function is $n \ln n$, while it is bound between $\sqrt{n}\ln(n)$ and $n\ln(n)$ for the UMDA. Asymptotic behavior of UMDA is better than that of SHC on the max A_{11} test case. Numerical experiments show that this asymptotic behavior does not take place until after $n = 200$ variables since SHC has always converged to the optimum faster than UMDA. It should be stressed that this study is primarily intended at characterizing the dimensionality effects on SHC and UMDA, irrespectively of other optimization difficulties. When multimodality is introduced through the separable function A_{66}, the reliability of the SHC is logically proportional to the percentage of the design space spanned by the basin of attraction of the global optimum while UMDA robustly finds the global optimum.

Bibliography

S. Droste, T. Jansen, and I. Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276(1-2):51–81, 2002.

J. Garnier and L. Kallel. Statistical distribution of the convergence time of evolutionary algorithms for long-path problems. *IEEE Transactions on evolutionary computation*, 4(1):16–30, 2000.

J. Garnier, L. Kallel, and M. Schoenauer. Rigorous hitting times for binary mutations. *Evolutionary Computation*, 7(2):173–203, 1999.

J. He and X. Yao. From an individual to a population: An analysis of the first hitting time of population-based evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 6(5):495–511, 2002.

T. Jansen and I. Wegener. On the utility of populations in evolutionary algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 1034–1041, San Francisco, California, USA, 7-11 2001. Morgan Kaufmann.

M. Mitchell, J.H. Holland, and S. Forrest. When will a genetic algorithm outperform hill climbing. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 51–58. Morgan Kaufmann Publishers, Inc., 1994.

H. Mühlenbein, T. Mahnig, and A.O. Rodrigez. Schemata, distributions and graphical models. *J. of Heuristics*, 5:215–247, 1999.

H. Mühlenbein and G. Paaß. From recombination of genes to the estimation of distributions, I. Binary parameters. *Lecture Notes in Computer Science 1141: Parallel Problem Solving from Nature IV*, pages 178–187, 1996.

M. Pelikan, D.E. Goldberg, and K. Sastry. Bayesian optimization algorithm, decision graphs, and Occam's razor. Technical Report 2000020, IlliGAL, May 2000.