Order statistics in artificial evolution

Stéphane Puechmorel, Daniel Delahaye

Order Statistics in Artificial Evolution

Stephane Puechmorel[1] and Daniel Delahaye[2]

[1] ENAC
7, Avenue Edouard Belin
31055 TOULOUSE CEDEX
[2] CENA

Abstract. This article deals with the exploitation of statistical information from extremes values of an evolutionary algorithm.
One can use the fact that upper order statistics of a sample converge to known distributions for improving efficiency of selection and crossover operators.
The work presented in this paper is restricted to criteria defined on real vector spaces. It relies on an underlying canonical model of genetic algorithm, namely tournament selection and uniform crossover. Nevertheless, the results obtained so far encourage further investigations.

1 Introduction

Evolutionary computation is recognized to be highly successful in solving real world optimization problems. Its robustness and ability to locate efficiently global optima among many local ones allows treatment of cases for which other methods fail. However, most of these properties rely on stochastic exploration of the search space and if the complexity of one iteration is low, the overall cost of the algorithm can be high due to the number of samples needed for locating the global optimum. Moreover, a compromise must be done between exploring the search space in order to find new candidate points and exploiting the already computed ones. Careful balancing is the key of speed-up in evolutionary computation and is a major area of research. The selection operator is the one which has greatest influence on this and several procedures has been designed, so that exploration or exploitation is enforced. There is two main classes of selection operators:

- Selection based on fitness. Operators belonging to this class use the value of the criterion to compute selection probabilities. The historical 'spin wheel selection' is one of them, like the 'stochastic remainder' which has better properties.
- Rank based selection. In this case, only the relative value of the criterion is used. All operator from this class may be thought as a two stage procedure: first sort the population, then draw with a given law inside the sorted sample. The popular tournament selection is rank based, and has a nice behavior on a broad range of problems. Furthermore, it is possible to tune the selection pressure by merely changing the size of the competitors pool.

In the following, we will restrict our attention to rank based selection, which will be the start point of our analysis (it is a quite consensual fact that rank based operators and specially tournament selection are among the best selection procedure available for evolutionary computation. Many of our own simulation agree with that point of view).

Another aspect of evolutionary algorithms is the design of mutation and crossover operators. Generally, mutation is an operator which avoids being trapped in local solutions and allows to enlarge the exploration scale. On the other hand, crossover are frequently used. In the following, we will address the problem of finding the global optimum of a criterion furnished by a real valued function f over an hyper-rectangle $\prod_{i=1}^{n}[a_i, b_i]$ so that operator will act on real vectors. Barycentric crossover is by far the most commonly used in this case. It is easy to see that the effect of crossing two parents is a uniform random trial on a segment containing them. The so called uniform crossover, that is drawing a new random linear combination for each component of the parent vectors is more efficient on many problems. Here again, it is easy to figure out that this operator samples points from a uniform deviation in the hyper-rectangle of $\mathbb{R}^n$ defined by the parents (the hyper-rectangle defined by two points x and y of $\mathbb{R}^n$ is the closed set $H(x, y) = \{u \in \mathbb{S}^n | \forall\, i = 1 \dots n, x_i \leq u_i \leq y_i\}$).

Gathering the previous notes, the process underlying a whole generation inside an evolutionary algorithm may be described by a sampling distribution based on mixed uniform deviates on hyper-rectangle and depending only on the previous population. This point of view has been adopted in [DMPJ01]. An evolutionary algorithm may then been seen has a measure valued dynamical system. Using asymptotic properties of upper order statistics, it is possible to compute a threshold above which samples may be considered as belonging to a neighborhood of a maximum, while samples below will be assumed to belong to an exploratory pool. The sampling distribution obtain at each generation will then be a mix of a uniform deviate on the complement set of the hyper-rectangle containing all extreme samples and an exploitation distribution defined on this hyper-rectangle.

Simulation results on standard test problems will be presented and a comparison done with the standard tournament selection - uniform crossover algorithm.

2 Asymptotic Law of Extremes

In this section we will collect some classical results about limiting distributions of order statistics. Fundamental references are [Zhi91], [dH81]. Let E be a compact of $\mathbb{R}^d$ and let $\{X_1, \dots X_N\}$ a size n sample of common distribution P on E. Finally, let $f : E \to \mathbb{R}$ a P-measurable function. For a measurable subset $U \subset E$, the essential supremum of f on U, denoted $essup f_U$, is defined as follows:

- Define the function $F_U : t \in \mathbb{R} \to P\left(\{u \in U | f(u) < t\}\right)$;
- $esssup f_U = \inf\{t | F_U(t) = 1\}$.

The essential supremum is the value that will be searched by a stochastic algorithm : high values taken by f on zero measure subsets will not be taken into

account. With the notations above, let M be the essential supremum of f on E and F be F_E. Put

$$V : x \in \mathbb{R}^+ \to 1 - F(M - x^{-1})\,;$$

V is said to be of regular variation at infinity if it exists a real number $\alpha > 0$ such that for all $t > 0$:

$$\lim_{x \to +\infty} \frac{V(tx)}{V(x)} = t^{-\alpha}\,.$$

The exponent α is called the tail index of the distribution F. When V is of regular variation, there exists a limiting law for the sample maximum $\eta_N = \sup\{X_1, \ldots X_N\}$ in the following sense. there exists real sequences $(c_n)_{n\in\mathbb{N}}, (d_n)_{n\in\mathbb{N}}$ such that:

$$\lim_{n \to +\infty} F^n(c_n x + d_n) = \psi_\alpha(x)\,.$$

With F the distribution function and ψ_α defined by

$$\psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha)\,, x < 0\,, \\ 1 \qquad\quad\;\; , x \geq 0\,. \end{cases}$$

One choice of such sequences is $d_n = M, \forall n$ and $c_n = F^{-1\leftarrow}(1 - n^{-1})$ (the notation $F^{-1\leftarrow}$ denotes the right limit). Many cumulative density functions are of regular variation. This is obviously the case if we can find $\alpha > 0, \beta > 0)$ such that :

$$F(x) = 1 - \beta(M - x)^\alpha + o((M - x)^\alpha)\,.$$

It may be noticed that the definition of regular variation can be extended by relaxing the assumption on the limit expression and imposing only that there exists an mapping $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that:

$$\lim_{x \to +\infty} \frac{V(tx)}{V(x)} = h(t)\,.$$

However, in this case we have for $(t_1, t_2) \in \mathbb{R}^{+2}$:

$$h(t_1 t_2) = \lim_{x \to +\infty} \frac{V(t_1 t_2 x)}{V(t_2 x)} \frac{V(t_2 x)}{V(x)} = h(t_1) h(t_2)\,.$$

It is a well known fact that continuous solutions of this functional equation are precisely of the for t^α, so that it turns out that the first form has full generality. Regular variation is not a stringent assumption. Most cumulative distribution functions obtained from maximization problems belong to this class. Furthermore, it can be shown [Zhi91] that $\alpha = d/2$ in the case of a class C^2 criterion with maxima inside E and uniform sampling in E. Following [Pic75], we define, for a threshold values ν, the conditional excess distribution function F_ν as:

$$F_\nu(t) = P(\{u \in E | \nu \leq f(u) < t\})/P(\{u \in E | f(u) \geq \nu\}) = \frac{F(t) - F(\nu)}{1 - F(\nu)},\; t > \nu\,.$$

If F_ν has regular variation of tail index α, putting $\nu = M - \theta^{-1}$ and $\lambda < 1$, one has

$$F_\nu(M - \lambda\theta^{-1}) = \frac{V(\theta) - V(\lambda^{-1}\theta)}{V(\theta)}\,,$$

so that

$$\lim_{\theta \to +\infty} F_\nu(M - \lambda\theta^{-1}) = 1 - \lambda^\alpha\,.$$

Except for the cases which can be analytically solved, the value of the tail index α can be estimated from the sample. Hill [Hil75] designed such a conditional maximum-likelihood estimate of α, which is easily computable:

$$\widehat{\alpha}_{k,N} = \left(\frac{\sum_{i=0}^{k-1} \log(\eta_{N-i})}{k} - log(\eta_{N-k}) \right)^{-1}$$

where $\eta_1 \geq \ldots \geq \eta_N$ are the order statistics of the sample and k is chosen from N so that $\lim_{N \to +\infty} kN^{-1} \to 1$.

3 Tournament Selection Analysis

3.1 Basic Results

Let n, m be integers such that $n \geq m \geq 1$. A (n, m)-tournament selection from a population $(X_i)_{i=1\ldots N}$ of size N produces individuals, first by uniformly sampling n individuals from the population (the pool of competitors), then selecting the m individuals with the highest (or lowest for minimization) criterion value. Increasing n increases the selection pressure and propension to elitism. Since pools of competitors are obtained by independent uniform sampling, noting will change if we apply a permutation σ to the population and apply tournament selection to the new population $(X_{\sigma(i)})_{i=1,\ldots,N}$. Among possible permutations, one may choose to order population in increasing order. Inside a pool of n individuals, those m with highest criterion value will be simply those with m highest indices. Formally, once the population has been order, one may describe the process of selecting m individual out of a pool of n by sampling the m highest order statistics from a discrete uniform law in the set $\{1, \ldots, N\}$, then selecting individual with matching indices. Join density of the m upper order statistics $\eta_{n-m+1}, \ldots, \eta_n$ of n independent uniform random variables is given by [Rei89]:

$$p_{\eta_{n-m+1},\ldots,\eta_n}(x_1, \ldots, x_m) = n! \frac{x_1^{n-m}}{(n-m)!}, \quad x_1 < \ldots < x_m\,.$$

For algorithmic implementation, it is convenient to realize sampling by successive draws from conditional laws. This can be done with the following procedure:

– Draw the real random variable y_m by sampling a real uniform random variable t in the interval $[0, 1]$, then computing $y_m = t^{n^{-1}}$.

– Assuming y_{k+1} has been obtained, y_k is obtained by sampling t and computing $y_k = y_{k+1} t^{(n-m+k)^{-1}}$.
– When all the y_i has been obtained, multiply each by N and round the result towards the nearest integer to obtain selected indices.

Some useful results may be obtained from this simple computation. First of all, probability of selecting the maximum of the sample increases nearly linearly with n in the case of large populations. In fact the probability of selecting the maximum in a population of size N and a $(n, 1)$-tournament with the rounding procedure described above is given by :

$$1 - \left(1 - \frac{1}{2N}\right)^n .$$

which can be expanded as

$$\frac{n}{2N} + o(N^{-1}) .$$

Second, in the case of a general (n, m)-tournament and if m is large (thus n), the selection of last individuals is close to uniform sampling (conditionally to the least index value obtained so far). A survey of properties of tournament selection may be found in [Bli97].

3.2 Limiting Distributions in Tournament Selection

As most evolutionary operators, selection may be viewed not as an operator evolving individuals but probability distributions [DMPJ01]. The most attractive feature about that is the ability to use classical convergence proof (fixed point theorems for example) instead of complicated probabilistic arguments. From that point of view, an evolutionary algorithm is a measure valued dynamical systems, which evolves empirical (that is to say sum of point distributions) measures. It may be noted that a weak law of large numbers exists within this frame, when the population size go to infinity. Our purpose is to restrict our attention to the tournament selection operator and to show how to use results on the asymptotic law of extremes. Let E be a Banach space (which will be $\mathbb{R}^n$ in our application), $\mathcal{B}(E)$ its Borel field and let $\mathcal{P}(E)$ the set of probability measures on E. The total variation distance on $\mathcal{P}(E)$ is defined by:

$$d(\mu, \nu) = \|\mu - \nu\| = \sup_{A \in \mathcal{B}(E)} |\mu(A) - \nu(A)| .$$

In the case of densities (with respect to the Haar measure on E), and confusing the notation of density and measure, we have the well known equality:

$$\|\mu - \nu\| = \frac{1}{2} \int_E |\mu(x) - \nu(x)| dx.$$

Now, let μ be a density and let $f : E \to \mathbb{R}$ be the criterion to be maximized that will be assumed of being twice continuously differentiable. The density of the best individual on a tournament of size n is given by:

$$n\mu(x)\mu\left(\{u | f(u) < f(x)\}\right)^{n-1} .$$

Let ϕ be the operator on probability measures associated with tournament selection. Taking densities μ, ν, we have:

$$\|\phi(\mu) - \phi(\nu)\| = \frac{1}{2}\int_E |\mu(x)\mu(L_{f(x)})^{n-1} - \nu(x)\nu(L_{f(x)})^{n-1}|dx$$

with $L_{f(x)} = \{u \in E | f(u) < f(x)\}$ the sub-level set of f on value $f(x)$. We can then write :

$$\begin{aligned}
&\|\phi(\mu) - \phi(\nu)\| \\
&\quad = \frac{1}{2}\int_E |(\mu(x) - \nu(x))\mu(L_{f(x)})^{n-1} - \nu(x)(\nu(L_{f(x)})^{n-1} - \mu(L_{f(x)})^{n-1})|dx \\
&\quad \leq \frac{1}{2}\int_E |\mu(x) - \nu(x)|dx + \frac{1}{2}\int_E \nu(x)|\nu(L_{f(x)})^{n-1} - \mu(L_{f(x)})^{n-1}|dx
\end{aligned}$$

Now, since$\mu(L_{f(x)})^{n-1} \leq 1$ and $\nu(L_{f(x)})^{n-1} \leq 1$, we get

$$|\nu(L_{f(x)})^{n-1} - \mu(L_{f(x)})^{n-1})| \leq (n-1)|\nu(L_{f(x)})^{n-1} - \mu(L_{f(x)})^{n-1}| \leq (n-1)\|\mu - \nu\|\,.$$

Finally

$$\|\phi(\mu) - \phi(\nu)\| \leq \frac{n+1}{2}\|\mu - \nu\|\,,$$

so that ϕ is Lipschitz.

From now on we assume that the set of critical points of f is finite and f reaches its maximum. Let $\mu \in \mathcal{P}(E)$; the $\phi\mu$-measure of $A \in \mathcal{B}(E)$ can be computed:

$$\begin{aligned}
\phi\nu(A) &= n\int_A \mu(x)\mu(L_{f(x)})^{n-1}dx \\
&= n\int_{-\infty}^{M} \mu(L_t)^{n-1}\int_{\partial L_t \cap A} \|f'(z)\|^{-1}\mu(z)dvol_{L_t}(z)dt\,,
\end{aligned}$$

where M is the maximum of f and $dvol_{L_t}$ is the canonical volume form on the level set L_t. To analyse further the behaviour of this measure, assume first that μ belongs to the domain of attraction of the weibull law, that is to say there exists a sequence (a_n) such that

$$\lim_{n\to+\infty} \mu(L_{M+a_n t})^n = e^{-(-t)^\alpha},\ t < 0\,,$$

and assume that f has a unique maximum at x_0, and is λ-convex. Let A be limited by a level set at the value t_0. Then, for n large enough, the $\phi\mu$ measure of A may be approximated by

$$I(A) = n\int_{t_0}^{M} e^{-\theta_{n-1}(M-t)^\alpha}\int_{\partial L_t \cap A} \|f'(z)\|^{-1}\mu(z)dvol_{L_t}(z)dt$$

with $\theta_n > 0$. Since f is C^2 and λ-convex, there exists $K > 0$ such that outside L_{t_0} those inequalities hold:

$$\lambda\|x - x_0\| \leq \|f'(x)\| \leq K\|x - x_0\|\,.$$

Therefore

$$K^{-1}n\int_{t_0}^{M} e^{-\theta_n(M-t)^\alpha}\int_{\partial L_t\cap A}\|z-x_0\|^{-1}\mu(z)dvol_{L_t}(z)dt \le I(A),$$
$$I(A) \le \lambda^{-1}n\int_{t_0}^{M} e^{-\theta_n(M-t)^\alpha}\int_{\partial L_t\cap A}\|z-x_0\|^{-1}\mu(z)dvol_{L_t}(z)dt\,.$$

4 Algorithm

Asymptotic distribution of order statistics allows some statistical inference about the value of the true maximum of a function given a sorted sample. Furthermore, rank based selection process like tournament naturally uses order statistics since relative value of individuals are used. To demonstrate the interest of statistical inference based on extreme values distribution, we have modified a standard tournament selection based on genetic algorithm so that sample values that may be considered as extremes will be specially treated.

4.1 Tail Index Estimation

Some runs of the standard GA has been done in order to test for adjustment of the distribution of population order statistics to a Weibull law.Since empirical distribution function can be easily computed from the sorted population the Kolmogorov-Smirnov test was used . For most generations, conformance of the upper order statistics to the asymptotic law is accepted. However, during some transition phases this property is lost and upper values no longer obey to a Weibull law. Hill tail index estimator yields values that have the same order of magnitude than the theoretical value for a C^2 criterion, but variation is high from a generation to another.

Maximum-likelihood estimation of both the tail index and the essential supremum of the criterion has been tried too, but increased computational cost (implicit equation solving) is a major drawback. Furthermore, tail index obtained by this procedure is close to the Hill estimator.

4.2 Genetic Operators on Extreme Values

It is easy to see that uniform real crossover is in fact uniform sampling in the hyper-rectangle defined by the two parents. That observation shows that some speed-up in convergence may be obtained by restricting the use of uniform crossover to individuals representing extreme values of the population. The only remaining point is to defined which values may be considered as extreme. The problem has been addressed in [GO03] but the solution found is computationally expensive (successive Kolmogorov-Smirnov goodness of fit tests). Following the same idea, we make adjustment tests but starting from the upper fifth distinct values of the population the increase by a fixed amount until KS test failed. We enforce that the number of tests be under some given number.

5 Results

Different test functions have been used in order to compare our method with a classical GA :

Sphere	$f_1(\boldsymbol{x}) = \sum_{i=1}^{i=N} x_i^2$	$-10000 \leq x_i \leq 10000$
Ackley	$f_2(\boldsymbol{x}) = -c_1.\exp(S_1(\boldsymbol{x})) - \exp(S_2(\boldsymbol{x})) + c_1 + e$ $S_1 = -c_2\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$ $with\ S_2 = \frac{1}{N}\sum_{i=1}^{N}\cos(c_3.x_i)$ $c_1 = 20\ c_2 = 0.2\ c_3 = 2\pi$	$-10000 \leq x_i \leq 10000$
Griewank	$f_3(\boldsymbol{x}) = \frac{1}{400*N}\sum_{i=1}^{N} x_i^2 - \prod_{i=1}^{N}\cos(\frac{x_i}{\sqrt{i}}) + 1$	$-10000 \leq x_i \leq 10000$
Rosenbrook	$f_4(\boldsymbol{x}) = \sum_{i=0}^{N-1} 100*(x_i^2 - x_{i+1})^2 + (1 - x_i)^2$	$-30 \leq x_i \leq 30$

All the functions have to be minimized and have their minimum at 0 unless the Lenard-Jones function for which only an experimental mininum is used (best known min=-128.287 for 30 atoms). It must be noticed that both algorithms use the same selection scheme (stochastic remainder without replacement which is not the best) and do not use any scaling or sharing operators.

Our goal being to compare the influence of domain chromosome and order statistics we wanted them to work exactly the same way from the selection point of view.

The number of evaluations being different at each generation for those two algorithms, the number of generations has been adapted in order to maintain the same number of evaluations for all experiments.

Notice that the following curves have been adjusted in order to represent both results on the same graph. Those adjustments have been done on both axes. The "x" axis address the number of evaluations for our GA and must be scaled for the standard GA ($\times$ 20). The "y" axis represent the fitness given by both algorithms. The given results are so different that a logarithm scale has been used to see both curves.

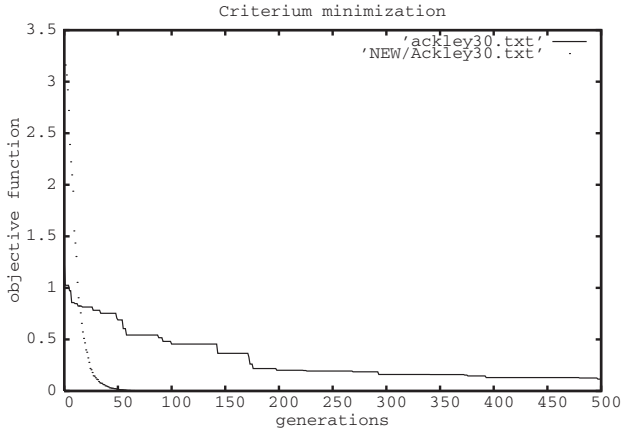
The parameters used for our GA are the following:

individuals 100	generations 500
probability of crossover 0.4	probability of mutation 0.3

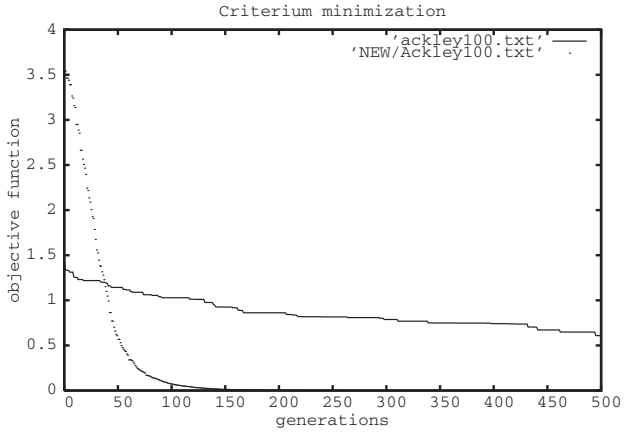
For the Rosenbrook, Lennard-Jones the number of generations has been extended to 1500 and 2500 respectively. The experiments have been done on a PentiumII 300 MHz and last 7 minutes for N=200 and 14 minutes for N=2000 (N is the dimension of the state space). It must be noticed that other experiments has been done for the same functions with the optimum moved in the state space (without symmetries) and the given results are quite similar.

Function	f_1	f_2	f_3	f_4
Standard AG - N=200	10621	11598	11.98	20.11
Domain AG - N=200	2.28	0	0.32	0.96
Standard AG - N=2000	910^5	8.710^5	20.45	165.8
Domain AG - N=2000	622	222	3.38	1.11

Function	f_5 N=200	f_6 N=90 (30 Atoms)
Standard AG	15.610^6	-77.21
Domain AG	254	-125.9 [1]

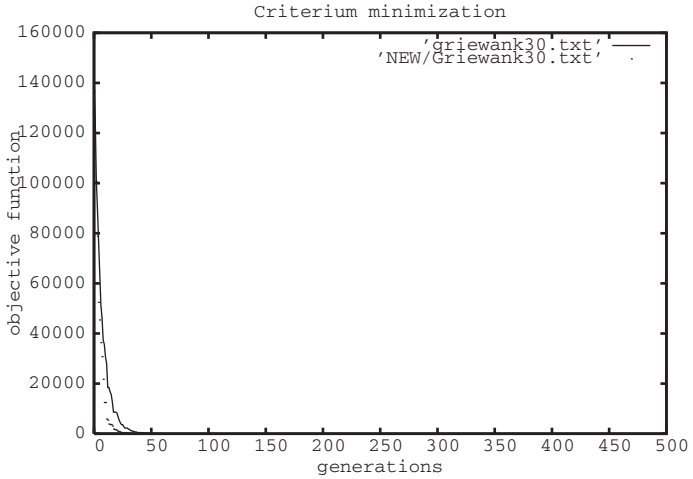


(a) Ackley Dimension 30

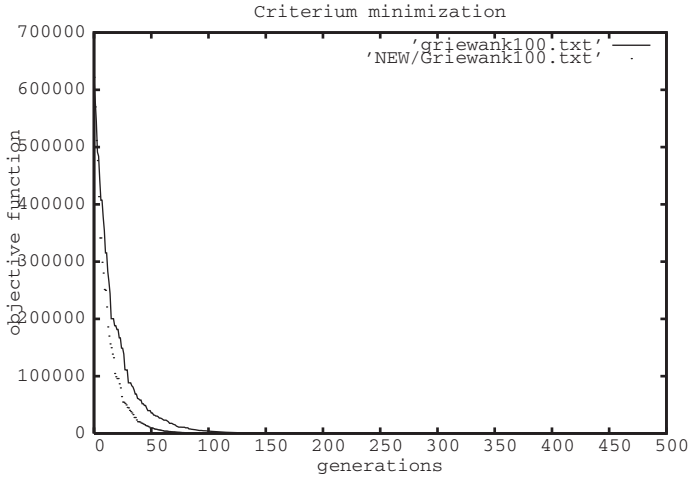


(b) Ackley Dimension 100

Fig. 1. Objective evolution for the Ackley function



(a) Griewank Dimension 30

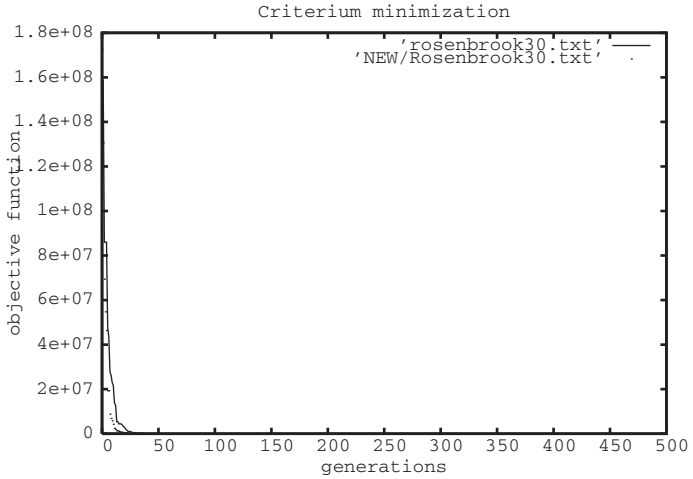


(b) Griewank Dimension 100

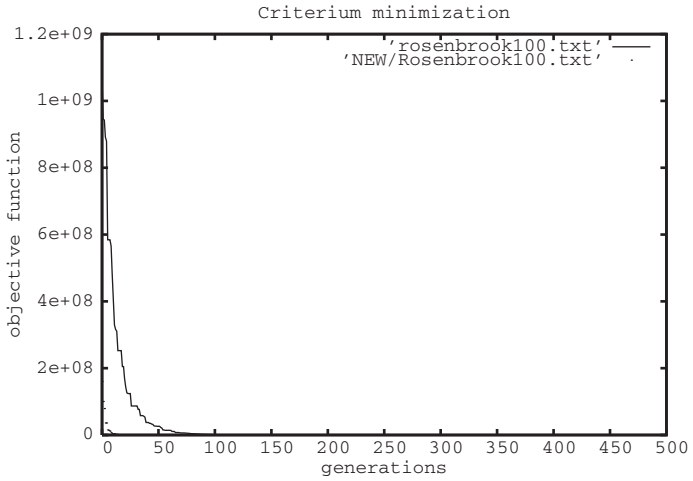
Fig. 2. Objective evolution for the Griewank function

6 Conclusion

This paper shows the gain given by the mix of extreme values inference and artificial evolution. On one side, the main advantage of order statistics for optimization is their abilities to summarize the properties of an entire domain with a "small" sample. On the other side, the evolution process of GA is able to build



(a) Rosenbrook Dimension 30



(b) Rosenbrook Dimension 100

Fig. 3. Objective evolution for the Rosenbrook function

the most adapted chromosome to environment given by the fitness landscape. The mixing of both methods really increase the performances of GA by guiding the exploration and exploitation phases. For all tests, results produced by this new GA, are better than those given by a classical GA.

Notice that this algorithm may be still improved in the following way:

- use of a better selection scheme;
- the order statistics may control the drawing of individuals;

– pools of samples may be stored to reduce the number of functions of evaluation.

References

[Bli97] T. Blickle. *Handbook of Evolutionary Computation*, chapter Tournament selection. IOP Publishing Ltd., 1997.

[dH81] L. de Haan. Estimation of the minimum of a function using order statistics. *Journal of the American Statistical Association*, 1981.

[DMPJ01] Kallel L. Del Moral P. and Rowe J. Modeling genetic algorithms with interacting particle systems. *Revista de Matematica, Teoria y aplicaciones*, 2001.

[GO03] J. GONZALO and J. OLMO. Which extreme values are really extremes. *Preprint*, 2003.

[Hil75] B.M. Hill. A simple approach to inference about the tail of a distribution. *Annals of Statistics*, 1975.

[Pic75] J. Pickhands. Statistical inference using extreme order statistics. *Annals of Statistics*, 1975.

[Rei89] R.D. Reiss. *Approximate distributions of order statistics*. Springer-Verlag, 1989.

[Zhi91] A.A. Zhighljavsky. *Theory of random search*. Kluwer Academic, 1991.