A New Stopping Criterion for Genetic Algorithms

Christelle Reynes and Robert Sabatier
Laboratoire de Physique Industrielle et Traitement de l'Information, EA 2415, Université Montpellier 1, 15 Avenue Charles Flahault, 34070 Montpellier, France

Keywords: Genetic Algorithm, Convergence, Stopping Criterion, Markov Chains, Simulation.

Abstract: Obtaining theoretically legitimate stopping criteria is a difficult task. Being able to use such criteria, especially in real-encoding context, remains an open problem. The proposed criterion is based on a Markov chain modelling and on the distribution of the number of occurrences of the locally best solution during several generations under the assumption of non-convergence. The algorithm stops when the probability of obtaining the observed number of occurrences is too small. The obtained criterion is able to fit very different solution spaces and fitness functions (within studied limitations) without any required user intervention.

1 INTRODUCTION

In a theoretical point of view, a Genetic Algorithm (GA) can be considered to have converged as soon as the global optimum is found. But in practical issues, convergence can only be detected by a persistence of an optimum for several generations and is rarely soundly adressed. In the proposed criterion, the number of generations without change of the current local optimum as well as the proportion of the population formed by the solution having the best fitness value will be taken into account.

Many studies arose about the design of a theoretical framework to assess GA convergence. The most important approach to model GAs is probably the use of Markov chains (Davis and Principe, 1993).

The scope of the proposed criterion could be linked to approaches such as takeover time and runtime modelling. However, the former is based on what happens without crossover and our objective is to model the whole behaviour of the GA. The latter concerns a much more theoretical framework than what is proposed here. Several studies have been proposed such as (Storch, 2008) but their main goal is to increase knowledge about algorithms behaviour and performances and not really practical applications.

Some stopping criteria have been proposed but most are based on binary encoding with rare extensions to alphabets whose cardinality is restricted to 2^k like in (Aytug and Koehler, 2000). The criterion proposed in this paper follows the lead of those researches as it also uses the Markov Chain formalism to derive its results by studying the expected behaviour of GAs.

The proposed criterion acknowledges a less rigourous theoretical framework but seeks for applicability to as many cases as possible regarding encoding strategies, operator use, fitness landscapes,... Of course, there are limitations which are clearly explained in the following sections. It is important to notice that this criterion does not claim to guarantee that the optimal solution has been found, that is why it will be called *pseudo-convergence* criterion. Obviously, for real optimization issues, it is impossible to ensure reaching the global optimum with heuristic methods but it is of big interest to have criteria to assess a good quality of the final solution. Here, everything will be done to calibrate the criterion so that it detects the convergence as quickly as possible but finding a good quality solution with more confidence will be favoured with regards to speed.

2 A NEW CRITERION FOR PSEUDO-CONVERGENCE

2.1 Overview of the Stopping Criterion

Our starting point is the following observation: a solution with a good fitness (locally or globally optimal) is likely to gradually overrun the population. This is due to selection, which favours survival of the best solutions, and is likely to be strengthened by elitism.

Reynes C. and Sabatier R..
A New Stopping Criterion for Genetic Algorithms.
DOI: 10.5220/0004154202020207
In *Proceedings of the 4th International Joint Conference on Computational Intelligence* (ECTA-2012), pages 202-207
ISBN: 978-989-8565-33-4

Our stopping criterion is based on the number of occurrences of the locally best solution (denoted LBS, that is to say the best one found so far) in the last populations. One occurrence is defined as one copy of the solution which currently achieves the best fitness value. As elitism is to be used, the LBS is obviously the best solution found so far.

The principle can be illustrated through a small simulated example (described in section 3.1). Fig. 1 shows the evolution of the number of occurrences of the LBS for 400 successive generations. After the line, the number of occurences of the LBS count occurrences of the global optimum whereas before the line, local optima were counted. It can be easily seen that the number of LBS significantly increases after this appearance.

However, the algorithm convergence cannot be questioned by considering only one generation. Indeed, the stochastic aspect of the algorithm involves constant fluctuations. Hence, the sum of the results of several successive generations will be used.

Conceptually, the criterion can be described as follows. The number of LBS occurences will be modeled for one generation (denoted S_1) and for the sum of w successive generations (denoted S_w, $w > 1$) **under the hypothesis that the global optimum has not yet been found**. After this modelling, it will be possible to associate a probability of obtaining an empirical value s_{obs}, $P(S_w = s_{obs})$, under this hypothesis. Thus, as the GA comes to convergence, the probability for S_w to take the observed value will become very small (let say less than p_{th}). Then, we will be able to consider that the underlying non convergence hypothesis is no more true and we will decide to stop the algorithm. Eventually, the criterion will be:

IF $P(S_w = s_{obs}) < p_{th}$ THEN stop the algorithm.

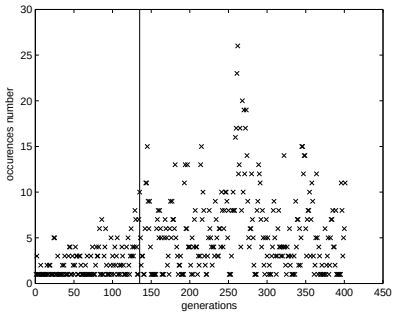

Figure 1: Evolution of the number of occurrences of the LBS for the 400 first generations for simulated data. The vertical line indicates the first appearance of the globally best solution in the population.

2.2 Definition of the GA Used

Real encoding will be used. Concerning selection, the fitness of the new solutions is computed and the solutions are ranked according to their fitness values (ties are averaged). Then, the selection probability for the r-th ranked individual is defined as $\mathbf{P}[\textit{select r-th ranked individual}\,] = \alpha \times r + \beta$.
where α and β are defined so that the sum over all the individuals of the selection probabilities is one and so that the probability to select the best individual is twice as high as the median ranked individual. Moreover, elitism is used: the best solution of the current population is automatically selected for the next generation.

In order to ease the modelling of the GA, mutation and crossover rates will be applied to individual solutions, and not on each encoding position.

2.3 Computation of $P(S_w = s_{obs})$

2.3.1 Modelling of the Number of LBS Occurrences for One Generation

Let $\{Z_n\}$ denote the process counting *the number of occurrences of the LBS in generation n*. Unfortunately, $\{Z_n\}$ does not only depend on $\{Z_{n-1}\}$ but also on the quality of other solutions constituting the previous population. Hence, in order to make it easier to use theoretical results, two hypotheses have to be assumed to consider $\{Z_n\}$ as a Markov chain.

Let $\{T_n\}$ denote a random variable which takes value 0 if the LBS has changed between generations $n-1$ and n (denoting that the GA has not converged) and value 1 if the same LBS has been kept. Then the first hypothesis required is the following one:

Hypothesis 1:
$\forall n \in \mathbb{N}, P(T_n = 1) = \varphi$ and $P(T_n = 0) = 1 - \varphi$, for some real constant $\varphi \in [0,1]$.

This hypothesis indicates that the probability of a local optimum change cannot be null (the global optimum has not been reached) and does NOT evolve along generations. In practice, this probability obviously changes. However, the most important is to obtain a modelling which is especially precise just before convergence. When the process is far from convergence, the model will over estimate the distance to convergence but the model will fit the process behaviour when the situation is decisive. That is why we will choose a value for φ which is close to 1 (see section 3.2 for more details).

Hypothesis 2: *The probability for new occurrences of the current LBS to appear from individuals which are not currently optimum is neglected.*

With this hypothesis we consider that only selection is responsible for increasing the frequency of the local optimum. We neglect the possibility for mutation and crossover to generate new occurrences of the currently considered LSB. This hypothesis will be of minor importance when the fitness function takes many different values and when the solution space is of high dimension.

Modelling:
Once these two hypotheses are assumed, Z_n value only depends on Z_{n-1} value and $\{Z_n\}$ can be considered as an order 1 Markov chain. Hence, its behaviour can be described through its initial state and transition probabilities, $\pi_n(k,l) = \mathbf{P}[Z_n = k|Z_{n-1} = l]$ (with $(k,l) \in \{1,2,\ldots,T_{pop}\}^2$, where T_{pop} is the population size). Two conditioning steps are required to compute these probabilities.

First Conditioning
$\pi_n(k,l)$ has to be split according to the two possible values of T_n. Z_n compulsory equals 1 if $T_n = 0$. Then, if $k = 1$, it may be due to a LBS change or the previous LBS may have been lost during mutation and crossover and retrieved thanks to elitism.

Second Conditioning
Let Z^{mc}_{n-1} denote the number of instances of the LBS remaining when mutation and crossover operators have been applied to generation $(n-1)$. If $Z_n = l$, according to the second hypothesis $Z^{mc}_n \in \{0,1,\ldots,l\}$. We obtain $\mathbf{P}[Z_n = k|Z_{n-1} = l, T_n = 1] = \sum_{j=1}^{l} \mathbf{P}[Z_n = k|Z_{n-1} = l, T_n = 1, Z^{mc}_{n-1} = j]\mathbf{P}[Z^{mc}_{n-1} = j]$.

To compute $\mathbf{P}[Z^{mc}_{n-1} = j]$, we have to consider p_m and p_c the respective probabilities of mutation and crossover for one solution. Then the probability for one solution to undergo at least one change is $p = p_m + p_c - p_m \times p_c$ (as probability and crossover are independent) and the probability to undergo no change is $q = 1 - p$. Finally, the distribution of Z^{mc}_n is given by a binomial one with parameters (l,q) :

$$\mathbf{P}[Z^{mc}_{n-1} = j] = \binom{l}{j} q^j (1-q)^{l-j} = b_{jlq} \qquad (1)$$

Now, elitism has to be taken into account as it adds one occurrence of the locally best solution. Hence, only $(k-1)$ occurrences have to be selected to obtain k occurrences in the next generation.
To compute the selection probability of the LBS, when it has j occurrences, we need to use the selection operator definition introduced in section 2.2. The rank affected to each occurrence of the LBS is $\overline{r} = T_{pop} - \frac{j-1}{2}$. Hence, the selection probability of each occurrence of the LBS is $p^*_j = \alpha\overline{r} + \beta$ and the selection probability of any of the j occurrences of the LBS is $j \times p^*_j$.
Outcome
Actually, here is the formula for $\pi_n(k,l)$ depending on the value of k:
if $k = 1$,

$$\pi_n(1,l) = 1 - \varphi\left[\left(\sum_{j=1}^{l} (1 - jp^*_j)^{T_{pop}} b_{jlq}\right) - 1\right] \qquad (2)$$

if $1 < k < T_{pop}$, $\pi_n(k,l) =$

$$\varphi\left[\sum_{j=1}^{l} \binom{T_{pop}}{k-1} (jp^*_j)^{k-1} (1 - jp^*_j)^{T_{pop}-k+1} b_{jlq}\right]$$

and if $k = T_{pop}$, $\pi_n(T_{pop},l) =$

$$\varphi\left[\sum_{j=1}^{l} \left(T_{pop}(jp^*_j)^{T_{pop}-1}(1 - jp^*_j) + (jp^*_j)^{T_{pop}}\right) b_{jlq}\right].$$

2.3.2 Modelling of the Number of LBS Instances along w Generations

Let define:

$$S^{(t)}_w = \sum_{i=1}^{w} Z_{t+i}.$$

Hypothesis 3:
We assume that $\{S^{(t)}_w\}$ is stationary, that is to say, its characteristics do not depend on time. In this case, it means that:

$$\mathbf{P}[S^{(u)}_w = j] = \mathbf{P}[S^{(v)}_w = j], \forall (u,v) \in \mathbb{N}^2,$$

It is then possible to simply study $S_w = \sum_{i=1}^{w} Z_i$.

Several simulation results (not shown here) showed that the stationarity hypothesis is not far from reality and takes into account that the locally best solution can still change which is completely consistent with the non convergence.
Now, to determine the distribution of S_w, we will firstly focus on the joint distribution of (S_w, Z_w) whose distribution is recursively assessed by

$$\forall w, \mathbf{P}[S_{w+1} = s, Z_{w+1} = k] = \sum_{l=1}^{T_{pop}} \pi_{w+1}(k,l)\mathbf{P}[S_w = s-k, Z_w = l]. \qquad (3)$$

To obtain S_w distribution, it is necessary to sum over all the states of Z_w.

2.3.3 Final Criterion

Actually, let us sum up the computation of $P(S_w = s_{obs})$:

$$\mathbf{P}[S_w = s] = \sum_{k=1}^{T_{pop}} \mathbf{P}[S_w = s, Z_w = k],$$

with

$$\forall w, \mathbf{P}[S_{w+1} = s, Z_{w+1} = k] = \sum_{l=1}^{T_{pop}} \pi_{w+1}(k,l)\mathbf{P}[S_w = s-k, Z_w = l]. \quad (4)$$

3 CRITERION STUDY

3.1 Threshold Determination

This illustration is a clustering problem. The goal is to conceive a GA that performs unsupervised learning with an unknown number of groups. Only a maximum allowed number of groups, K_{max}, has to be a priori given. The approach chosen here is to optimize the assignment of the observations to groups, making the issue a combinatorial problem.

During initialization, for each potential solution in the population, a number of groups, k, is uniformly randomly chosen in $\{2,\ldots,K_{max}\}$. Then, an integer in $\{1,\ldots,k\}$ is uniformly randomly chosen for each of the n observations. Concerning mutation, three possibilities are allowed: withdrawing or adding a group and changing one or more assignment(s). The crossover is a uniform one. Finally, the fitness function depends on the number of groups, k, and on the sum of within-group variances, in order to take into account both parsimony and precision of the model.

The dataset is a simulated one, so that the global optimal solution is known. The data contain 80 observations divided into four groups and described by five features. The first two features give the location of the observations in a plan whereas the other three ones are only uniform noise. For each group, the values of the first two variables are generated by a normal distribution whose average gives the centroid location and whose variance indicates the range.

To perform a first, coarse estimation of the satisfying p_{th}, the described GA has been applied six times on this dataset and we observed the evolution of S_{20}. For these runs, the global optimum has been found by the GA after respectively 109, 77, 71, 98, 70 and 73 generations. If we choose $p_{th} = 10^{-3}$, the global optimum is missed in three out of six runs (results not shown). For $p_{th} = 10^{-4}$, one more run is successful. From $p_{th} = 10^{-5}$, the global optimum is always found.

In order to test $p_{th} = 10^{-5}$, the GA was run one hundred times leading to 91% of optimum discovery (with only one misclassification in the remaining 8% and two in the last 1%). It would be possible to choose a smaller threshold. However, the efficiency improvement would be small whereas the computation time before stopping would be much lengthened. That is why we chose $p_{th} = 10^{-5}$.

By applying the formula obtained in section 2.3 with the chosen parameters, Tab. 1 gives some minimum values of s_{obs} required to stop the algorithm for usual values of p_m, p_c and T_{pop}.

3.2 Parameters Influence

The criterion definition implies that it depends on GA parameters but also on its window size w, $\mathbf{P}[T_n = 0]$ and the threshold previously studied. The continuation of that section will deal with the study of the first two parameters influence. This can be done regardless of the optimization problem which does not interfere in the distribution computation.

$\mathbf{P}[T_n = 0]$ has to be constant due to the 1^{st} hypothesis. We have determined $\mathbf{P}[T_n = 1] < 1$. Moreover along generations, $\mathbf{P}[T_n = 0]$ will rapidly become very weak. Thus, small values are going to be studied.

The results can be found in Fig. 2. The plotted value roughly shows the proportion that must be filled up by the LBS before deciding to stop the GA. As expected, the less the probability to find a better solution, the less this proportion. So that the criterion is more stringent, small values of $\mathbf{P}[T_n = 0]$ will be favoured. From now on, we choose $\mathbf{P}[T_n = 0] = 0.01$.

Concerning the window size, Fig. 2 shows values between 3 and 50. For small w the filled up proportion has to be more important and rapidly decreases when w value increases. From twenty generations, the decrease slows down that is why we chose a value of $w = 20$ for further applications.

3.3 Limitations

The 1^{st} hypothesis in section 2.3.1 has been studied in previous paragraphs. Even if it is a strong hypothesis, taking a small value for $\mathbf{P}[T_n = 0]$ allows to minimize the consequences with regards to convergence. The objective of this section is to highlight cases for which the second hypothesis cannot be assumed.

Firstly, if the considered issue deals with a solution space whose dimension is small, the probability to obtain several times the same solution cannot be

Table 1: Minimum values of s_{obs} required to stop the algorithm for $p_{th} = 10^{-5}$, $w = 20$, $p_m \in \{0.5, 0.6, 0.7\}$, $p_c \in \{0.5, 0.6, 0.7\}$ and $T_{pop} \in \{50, 100, 200, 500\}$.

		p_c = 0.5				p_c = 0.6				p_c = 0.7			
		50	100	200	500	50	100	200	500	50	100	200	500
p_m	0.5	126	130	132	134	97	99	100	101	76	77	78	78
	0.6	97	99	100	101	80	81	82	82	66	67	67	67
	0.7	76	77	78	78	66	67	67	67	57	57	58	58

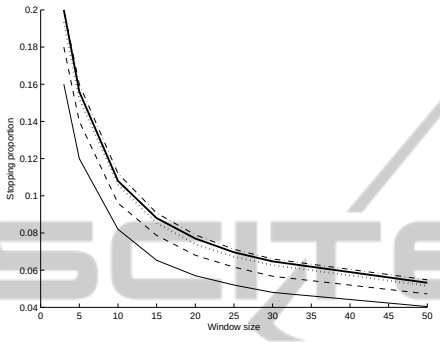

Figure 2: Evolution of the stopping criterion with the window size and $\mathbf{P}[T_n = 0]$. The solid line curve represents $\mathbf{P}[T_n = 0] = 0.5$, the dashed one to $\mathbf{P}[T_n = 0] = 0.25$, the dotted one to $\mathbf{P}[T_n = 0] = 0.1$, the bold one to $\mathbf{P}[T_n = 0] = 0.05$ and the dot-dashed one to $\mathbf{P}[T_n = 0] = 0.01$.

neglected. However, the optimization in such spaces is quite easy and does not require the use of a GA. Yet, the same phenomenon would occur for many solutions having the same fitness value. For instance, this can happen if the fitness function is a misclassification with few individuals to be classified. In such a case, if the solutions are really equivalent in the application context, the problem is likely to be solved by a simpler optimization method, else the fitness has to be reformulated to take into account this variety.

On the other hand, when initialization is not completely random but focused around chosen points our criterion should not be used.

4 APPLICATIONS

The first application will allow a detailed study of parameters influence including function complexity and GA parameters. Finally, some usual test problems for optimization will be used to show the results of the criterion for various and difficult functions.

4.1 Rastrigin's Function

The generalized Rastrigin function (Mühlenbein et al., 1991) is a usual non linear multimodal function used to test optimization methods. This function presents many close local minima and only one global minimum. The shape of the function is determined by the external variables A and L, which control the amplitude and frequency modulations respectively. The global minimum is 0.

Concerning Rastrigin's parameters, for A, integer values between 2 (the amplitude in the data is then about 55) and 15 (the amplitude is about 100) will be considered. The effect of L is really important, for $L = 1$, the solution space contains 25 local minima and for $L = 5$, we find 729 minima. Thus, integer values between 1 and 5 will be considered.

For the first simulation, we studied the influence of A and L for $p_m = 0.6$, $p_c = 0.5$, $T_{pop} = 100$ and $d = 2$. For each combination between A and L, fifty runs of the GA have been performed.

For all the runs, the final solution was into the deepest hole of the function even if the average number of generations required to see the global optimum for the first time increases with both values of A and L.

Then, the GA parameters, $T_{pop} \in \{50, 100, 200\}$, p_c and p_m $in\{0.5, 0.6, 0.7\}$ are studied for $A \in \{2, 25\}$ and $L \in \{2, 5\}$. For each combination, 450 runs have been performed. Excepted the case $A = 15$ and $L = 5$, the error rate is very low (0% in 59% of combinations, less than 1% of errors in the remaining combinations) and seems to be independent on the GA parameters. In complex issues, the stopping criterion is slightly less efficient (but it only fails in maximum one out of ten trials) and requires appropriate GA parameters. Hence, our stopping criterion is really helpful but does not make it any the less necessary to look for appropriate parameters for the most complex problems.

The last studied parameter is the Rastrigin dimension d. Eight values have been chosen between 3 and 10. For the easiest combinations, the stopping criterion performs very well (at most, the deepest hole was missed twice). For $5 \leq d \leq 9$ the optimum is missed in at most 20% of runs but for $d = 10$, 17 out of 50 runs were not successful. These cases correspond to really complex situations and it is not really surprising to miss the real optimum for certain trials. Here, it can be interesting to notice that, in a general point of

Table 2: Convergence results. In the first three columns, range, average and standard error (between brackets) values of the objective functions are given. The last column indicates the theoretical optimum value.

function	y_{min}	y_{max}	$\bar{y}$	y_{opt}
Osborne	6e-5	5e-3	2.54e-4 (5e-4)	5.46e-5
Bard	8.2e-3	8.2e-3	8.2e-3 (4e-7)	0.008215
Biggs	1.2e-6	5.5e-3	5.26e-4 (1e-3)	0
Gulf	8.4e-32	8.2e-5	2.2e-6 (6e-6)	0

view, it is always reasonable to perform a GA several times to evaluate the solution robustness.

4.2 Application to Standard Test Problems

A subset of test functions in (Moré et al., 1981) consisting in sum of squares of n_f functions of n_v variables is used: namely Osborne I, Bard, Biggs EXP6 and Gulf Research and Development. Results are introduced in Tab. 2.

In all cases, solutions obtaining very good values of the objective functions have been found during the different runs and the worst objective function value obtained is always close compared to the real range of it. Hence, running the GA a few times (which can be considered as compulsory when dealing with stochastic optimizers) using the proposed stopping criterion is likely to bring much information about the true solution. When no information is known about the objective function behaviour, it could be really difficult to decide to stop after any given number of generations. Indeed, considering these functions, the criterion required between a few hundreds and several tens of thousands of generations to stop.

5 CONCLUSIONS

Thanks to the modelling of the process describing the number of occurrences of the LBS during several successive generations, a new stopping criterion has been proposed for real-encoded GAs. The originality of our criterion is on one side the focus made on the LBS occurences and on the other side, the generality of its use: operators are completely free as long as they respect the definition of the mutation and crossover rates and especially the criterion has been developed to apply on real-encoded GAs. It has the main advantage of taking into account all the GA operators without requiring user intervention when changing problem. The modelling required three hypotheses implying some cases where this stopping criterion should not be applied.

Despite the required simplifications, the theoretical developments performed in this paper allow to provide a useful understanding of GA unfolding even if they do not restore the whole complexity of reality. This distance between the model and the real situation leads us to consider a very small probability (10^{-5}) for the algorithm stopping. In our opinion, this distance is mainly due to the second hypothesis.
Concerning the first hypothesis, the most stringent case has been chosen. Then, we probably would be able to stop earlier without missing the global optimum. However, the main goal of this criterion is not to achieve speed performances. It is more specifically designed to enable the user to obtain a good solution without intervention in the GA stopping process.

Actually, even if the model does not perfectly fits, the simulations performed in this paper proved the stopping criterion efficiency. Our stopping rule appeared to be equally efficient for completely different and very complex functions. Robustness was also shown concerning changes in the GA parameters.

Actually, the proposed stopping criterion should be used instead of arbitrary criteria, for problems within limitations of section 3.3. It does obviously not guarantee to find the global optimum, hence the GA has to be run several times.

REFERENCES

Aytug, H. and Koehler, G. (2000). New stopping criterion for genetic algorithms. *European Journal of Operational Research*, 126(3):662–674.

Davis, T. and Principe, J. (1993). A markov chain framework for the simple genetic algorithm. *Evolutionary computation*, 1(3):269–288.

Moré, J., Garbow, B., and Hillstrom, K. (1981). Testing unconstrained optimization software. *ACM Transactions on Mathematical Software (TOMS)*, 7(1):17–41.

Mühlenbein, H., Schomisch, M., and Born, J. (1991). The parallel genetic algorithm as function optimizer. *Parallel computing*, 17(6-7):619–632.

Storch, T. (2008). On the choice of the parent population size*. *Evolutionary Computation*, 16(4):557–578.