THE POPULATION SIZING PROBLEM IN (P)GAS

Zdeněk Konfršt
Department of Cybernetics, Czech Technical University

Abstract

One of interesting problems in genetic algorithms (GAs) is the population sizing problem. If you take a population too small, the final solution (represented by the quality of the whole population) is suboptimal and therefore the wrong one. On the other hand, if you take the population too large, you must reach the right target but you lose a computing time on processing redundant individuals. In order to estimate the correct size of population, approximations of population sizing are used. They consider an approximated function, a representation of individuals and the character of used operators. Computing environment of Matlab allows us, with help of the estimation model, to identify the right population size, to estimate the final overall quality of individuals and to compare theoretical results with real runs of GAs.

1 Introduction

The genetic algorithms are stochastic based algorithms which are influenced by various parameters. The most important ones seem to be the size of population, the crossover scheme, the probabilities of crossover and mutation, the selection scheme and the replacement strategies. Some models of simple genetic algorithm (also SGA) have been created [4] however, they still do not give all answers to all questions we are facing in the area of GAs.

As it seems natural that the description of GA[1] run is quite simple, the mathematical descriptions may lack in many views. The truth is that the number of models increases using the theories of statistics and probability. In the article, we follow the estimation theories of the expected quality based on a population size. We choose the most mature one, which employs a well-know problem—one dimensional random walk (sometimes called gamler's ruin problem). We clear up the basics and carry out some appealing experiments.

The article is organized into several sections. Introduction is the first one. Second one, Backgroud reviews the published theory on population sizing and concentrates on the estimation based on one dimensional random walk. Third section (Domains of interest) explains tasks we tackle. In Experiments, results of experiments are shown mainly through figures and a table. The Discussion section explains, clarifies the results and kicks off points for further work. Conclusion, the last one, summarizes the results.

2 Background

This section is a review of the most important work on population sizing and its relation to the convergence quality of GAs. We deal with fundamental milestones in several subsections. In Decomposition, key findings about GAs are presented, which have become the "golden" rules in the area of GAs over the years. In De Jong's estimate, a first population sizing equation is stated and the linkage with 2^k(two)-arm bandit problem is mentioned. The next section-Conservative estimate presents a first full-grown model. In the last section, the newest model is presented and fully described.

2.1 Decomposition

In GAs, chromosomes, often represented as binary vectors, encode the solution. Combination of bits (schemata) acts as a unit to influence the fitness of an individual is known as building

---

[1] In the article, we employ the term "GA" instead of "PGA of one population". PGA of one population belongs to the main scopes of our study. As it appeared in [10], those two types of GAs could be easily transformed between each other.

blocks (BBs). In [5], it was firstly presented a sort of decomposition of GA based upon BBs. The highlights are reviewed:

- The GA is processing BBs.
- Ensure the growth of necessary.
- Supply enough BBs in the initial population.
- Mix the BBs properly.
- Solve problems that are BBs tractable.
- Decide well among competing BBs.

Those were pointed out as the basics on the BB theory. We do not inquire them again, because it could be found elsewhere [5,6,8].

2.2 De Jong's estimate

Holland [1,2] idealized the process in a GA as a cluster of parallel and interconnected 2^k-armed bandit problems. The Holland's bandit problem was extended by De Jong [3] in the equation for population sizing as

$$n \geq \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{|f_1 - f_2|} \tag{1}$$

where n is the size of population, σ_1^2 and σ_2^2 are the variance values of the two arm payoffs respectively,f_1 and f_2 are the mean fitness values. It is representing the basic noise-to-signal equation.

Although it was not extended in the De Jong's dissertation, it gave a first approximation on the population sizing problem.

2.3 Conservative estimate

In [5], the statistical decision theory was employed. They have been modelling a GA run as competitions between the best and the second best BBs. Those two BBs were represented with their mean fitnesses and fitness variances. One of the outputs was that the probability of the right choice in a single trail in a problem (with m equally sized partitions) is

$$p = \Phi(\frac{d}{\sqrt{2m'}\sigma_{bb}}) \tag{2}$$

where Φ is the cumulative distribution function (CDF), σ_{bb} is the average BB variance, m' is a number of competing partitions $(m-1)$ and d is the fitness difference between the best and the second best BBs. The equation below gives the population sizing estimate as

$$N = 2c(\alpha)m'(1+\rho_t^2)\chi^k\frac{\sigma_M^2}{d^2} \tag{3}$$

where n is the population size, $c(\alpha)$ is the probability of a good solution, m' is a number of competing partitions, ρ_t^2 represents additive noise[2], σ_M^2 is the average fitness variance of the partition being considered, d^2 represents second power of the fitness difference between the best and the second best BBs and χ^k could be substituted as 2^k, because normally every bit in a string could contain just two values $0, 1$. The variable k equals 1 when the number of partitions equals the length of a string.

The approximation (3) stated above gives even a softer solution to De Jong's one on population sizing.

---

[2] Without noise the term $(1+\rho_t^2)$ is omitted.

2.4 The Gambler's ruin problem

In [8], they used a well known one-dimensional random walk with absorbing bariers $x = 0$ and $x = n$ representing convergence to the wrong and the right solution, respectively. Probabilities p and $1-p$ are representing probabilities that the best BB takes over the second best or vice versa. They have been using an initial seed defined as $x_0 = \frac{n}{2^k}$ (of k order) and assuming that the competition takes place between the best and the second best BBs in a partition, crossover and mutation do not destroy significant numbers of BBs, boundaries of the random walk are absorbing [3].

The well-known equation[4] [7] was employed to get the quality of a solution as the number of X partitions converged to the right BBs.

$$P_{bb} = \frac{1 - (\frac{1-p}{p})^{n/2^k}}{1 - (\frac{1-p}{p})^n} \tag{4}$$

where P_{bb} is the quality of a final population, the initial population is $x_0 = n/2^k$, p is the probability that the best BB takes over the second best and it should hold $p > 1-p$. We get p from the equation (2).

The size of population was defined as

$$n = -2^{k-1} ln(\alpha) \frac{\sigma_{bb}\sqrt{\pi m'}}{d} \tag{5}$$

where n is a population size, α is the term of premature convergence ($\alpha \leq 1$, often $\alpha = 0.01$ or $1/L$, where L is the length of a string) and k is a maximum number of independent variables (it holds $L = k.m$). This equation (5) shows the relations among the variables and we could summarize:

- Long BBs (large k) are more difficult to solve because longer BBs are scarcer in the initial population.
- The required population is inversely proportional to the signal-to-noise ratio.
- Longer problems (large m) are more difficult because there are more sources of noise than in short ones.

3 Domains of interest

In the previous section, the population sizing problem has been explained. The last estimation model is our primary concern. We would like to answer the items below:

- How does the characteristic $p = \Phi(d, m')$ look in a 3-D graph?
- Show the characteristic $P_{bb}(n)$ based on variant p and k parameters.
- Compare $P_{bb_{theor}}$ and $P_{bb_{exper}}$ (with and without a mutation operator).

4 Experiments

In Figure 1, the equatation (2) is displayed for d and m' as parameters in a 3-D graph[5].

---

[3] This is not completely true while using mutation. That operator could destroy an "already absorbed" individual. It may have been one of the reasons why they did not use a mutation operator in their experiments at all.

[4] A result from the theory of random walk is that a particle will be captured by the absorbing barrier at $x = n$.

[5] Slicing the graph for either $d = const$ or $m' = const$, we get desired 2-D characteristics.

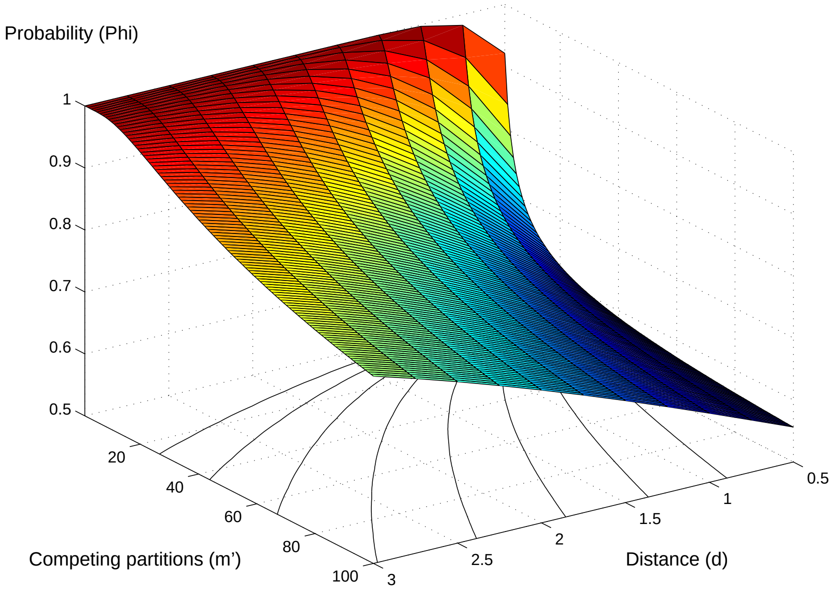

Figure 1: $p = \Phi(d, m')$

In Figure 2 (a left image), the equation (4) is shown for various probabilities p, $\sigma = 0.25$ and $k = 1$. The probabilities are marked in the graph according to a legend. For example, the probability 0.501 is stated as $p501$ and similarly. The right image shares the legend with the left one. The input values differ just in k ($k = 6$).

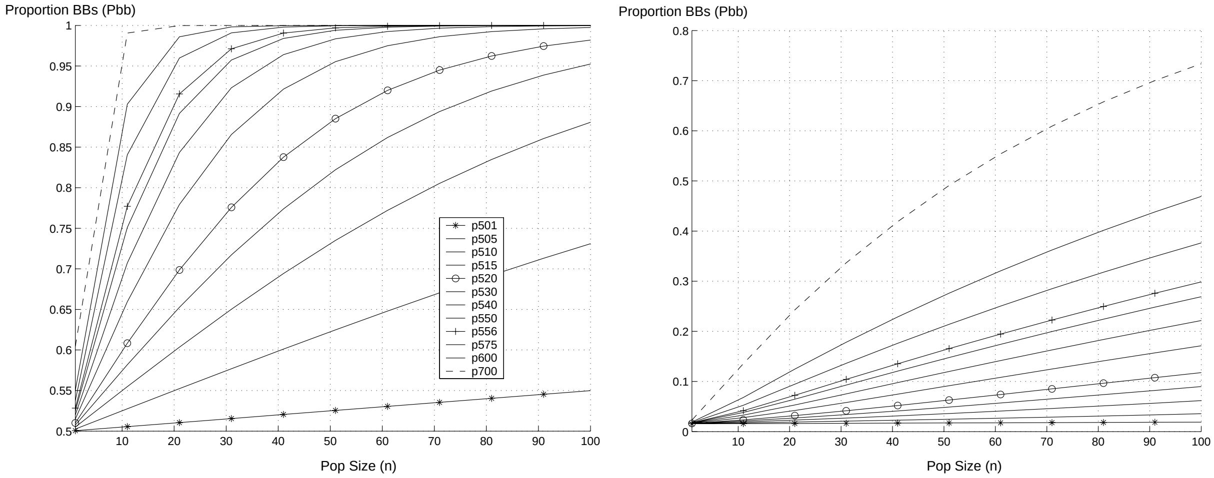

Figure 2: $P_{bb}(p, n, k)$

Concerning experiments, we used uniform crossover with p_c probability, tournament selection p_{ts}, generational replacement scheme, binary data type, the string length $L = 100$, mutation with probability p_m and the test function—Onemax. The results are average of 10 independent runs with random intialization and the stopping condition—"too similar" with the value 0.99.

In Table 1, there are theoretical and experimental values of $P_{bb}(n)$. The values are characterized by the vector $< dataX, p_c, p_m, p_{ts} >$, where the label $dataX$ (X is a number) distinguishes a curve in Figure 3. Note the only first row contains theoretical values and the first row is characterized by the first vector and so on in a vector list. So the order matters and the vectors are $< 5, 0.5, 0.0, 0.5 >$, $< 1, 0.6, 0.0, 0.6 >$,$< 2, 0.5, 0.0, 0.5 >$,$< 3, 0.5, 0.01, 0.5 >$ and $< 4, 0.5, 0.001, 0.5 >$.

In Figure 3, there are shown curves, which are based on the data from Table 1. Concerning

n	2	10	20	30	40	50	60	70	80	90	100
P_{bb}	0.5565	0.7770	0.9155	0.9712	0.9906	0.9969	1.0	1.0	1.0	1.0	1.0
P_{bb}	0.53	0.63	0.89	0.94	0.97	0.98	0.99	1.0	1.0	1.0	1.0
P_{bb}	0.53	0.65	0.87	0.93	0.96	0.95	1.0	1.0	1.0	1.0	1.0
P_{bb}	0.70	0.90	0.94	0.96	0.96	0.97	0.97	0.97	0.98	0.98	0.98
P_{bb}	0.57	0.76	0.90	0.95	0.99	0.99	0.99	1.0	1.0	1.0	1.0

Table 1: Data $P_{bb}(n)$ based on the theory and experiments.

the legend, the interpolation of $dataX$ is marked as $dataXi$.

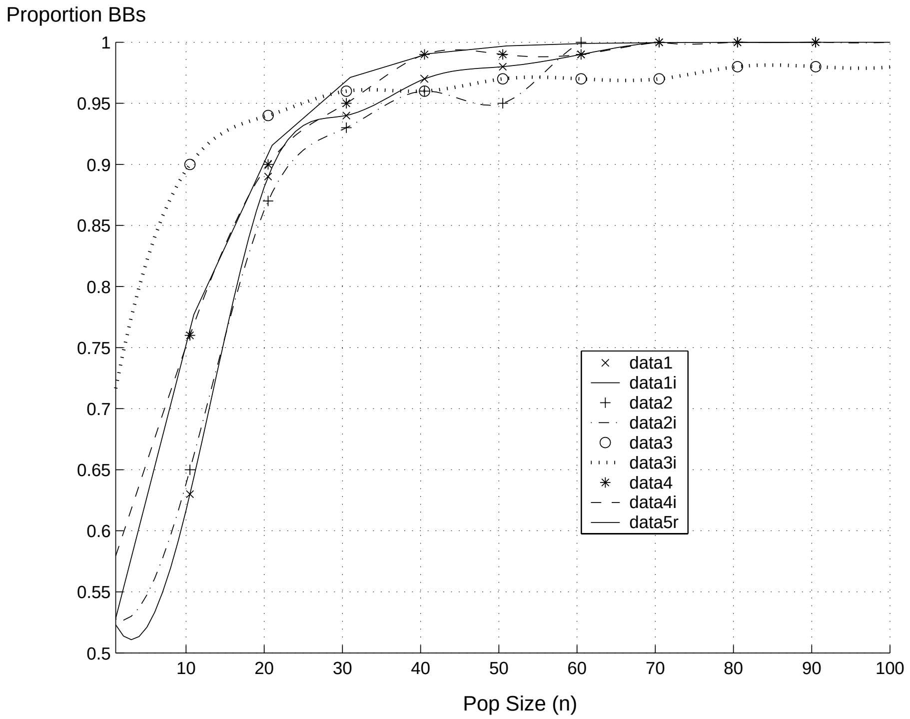

Figure 3: Experimental runs of GAs and their interpolations.

4.1 Discussion

As it has been shown in Figure 2, the variable k is an important parameter because a it grows it quickly disables the ability of a GA to reach a sufficient proportion of BBs. In those cases, higher probabilities of p are necessary.

As real runs of GAs and the model of GAs are concerned, they match fairly under one important condition. It is that zero or very low ($p_m < 0.001$) mutation is used. The estimation model was designed in a such way that the only disrupting process is a random generation of individuals in the beginning.

Some new appealing points have appeared as to carry out similar experiments for multi-deme PGAs and to undertake a better study with more parameters of GAs and with a wider range of test functions involved.

For the theoretical approximations, we have used Matlab[6] 6.0 (Release 12) running on IBM SP2 (spe103 node) with AIX OS (4.3.3). To get a hint of behaviour of real GAs we modified routines from PGAPack Parallel Algorithm Library (1.0) from the Argonne National Laboratory. The computing machinery is located at the Supercomputing Center, CTU, Zikova 4, Prague 6.

---

[6] The product of Mathworks, Inc. *http://www.mathworks.com*

5 Conclusion

The population sizing problem of GAs has been explained and the relevant theory of population sizing was reviewed. The main stress was put on examining the population estimation model based on one-dimensional random walk. Some compelling questions were raised and carried out experiments gave appropriate and acceptable answers. Some issues induced from experiments were discussed.

Acknowledgments

The research has been supported by the Ministry of Education of the Czech Republic within the Project No. MSM212300014: Research in the Area of Information Technology and Communication and the Czech Technical University Grant No. 300008203: Parallelization of Genetic Algorithms.

Contact

The author can be reached at `konfrst@zikova.cvut.cz` or at Department of Cybernetics, Czech Technical University, Karlovo náměstí 13, 121 35 Prague 2.

References

[1] Holland J. H.: Genetic algorithms and the optimal allocations of trials. *SIAM Journal of Computing, 2*(2), 1973, pp. 88-105.

[2] Holland J. H.: *Adaptation in natural and artificial systems.* Ann Arbor, MI: University of Michigan Press, 1975.

[3] De Jong K. A.: *An analysis of the behaviour of a class of genetic adaptive systems.* Doctoral dissertation, University of Michigan, Ann Arbor. *(University Microfilms No.76-9381)*, 1975.

[4] Vose M.: Modeling simple genetic algorithms *Foundations of Genetic Algorithms, II.*, 1992, pp. 63-74.

[5] Goldberg D. E., Deb K., Clark J. H.: Genetic algorithms, noise, and the sizing of populations. *Complex Systems, 6.*, 1992, pp. 333-362.

[6] Goldberg D. E., Deb K., Thierens D.: Toward a better understanding of mixing in genetic algorithms. *Journal of the Society of Instrument and Control Engineers, 32* (1), 1993, pp. 10-16.

[7] Feller W.: *An introduction to probability theory and it's applications* (2nd ed.), Volume 1, Wiley, 1966.

[8] Harik, G., Cantú-Paz E., Goldberg D. E., Miller, B. L.: The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Proceedings of the IEEE International Conference on Evolutionary Computation*, 1997, pp. 118-125.

[9] Deb K., Agrawal S.: Understanding interactions among genetic algorithm parameters. *Foundations of Genetic Algorithms, V.*, 1999, pp. 265-286.

[10] Konfršt Z.: *Parallel Genetic Algorithms.* Gerstner Laboratory Report 82/99, CTU Prague, 1999.