Population Size and Sampling Complexity in Genetic Algorithms

Yong Gao
Department of Computing Science
University of Alberta
Edmonton, Canada T6G 2E8
ygao@cs.ualberta.ca

Abstract

Determining an appropriate population size is an important task in the design of genetic algorithms and is closely related to the principle of implicit parallelism. In this paper, the problem of bounding the population size is formulated as that of minimization of sampling errors. Using techniques from computational learning theory, two lower bounds on the population size are established. The lower bounds enable us to interpret the principle of implicit parallelism from a new perspective, and depict how the necessary population size is related to the mutation probability and some population statistics.

1 INTRODUCTION

Determining an appropriate population size is an important task in the design of genetic algorithms and is closely related to the principle of implicit parallelism. Based on Holland's schema theorem, the principle of implicit parallelism is usually understood as"the number of schemata processed effectively is proportional to the cube of the population size n". In fact, different orders of estimations could be obtained under different assumptions on the relations between the population size N and the encoding length. Further discussion on the derivation of the principle of implicit parallelism can be found in [Bertoni and Dorigo, 1993].

In the practice of designing efficient genetic algorithms, there has been strong empirical evidence showing that population size is one of the most important parameters that plays a significant role in the performance of the genetic algorithms [Leung *et al.*, 1997, Grefenstette, 1986, Eiben *et al.*, 1999]. There also has been some theoretical work dealing with the problem of bounding the population size. For example, in [Goldberg *et al.*, 1992, Harik *et al.*, 1997] population size equations were derived by considering a specific set of competing schemata under some assumptions about statistical independence. In recent years, much attention has been paid to the design of sampling-based genetic algorithms [Leung *et al.*, 2001, Muhlenbein and Mahnig, 1999, Pelikan *et al.*, 1999]. From the statistics and machine learning point of view, an appropriate population size is even more critical to the success of this type of algorithms.

In this paper, we show that the problems of bounding the population size can be studied by using techniques from the computational learning theory. By formulating the genetic algorithm as an algorithm that explores the search space by sampling, two lower bounds on the population size are established under certain sampling error criteria. The lower bounds enable us to understand the principle of implicit parallelism from a new point of view, and relate the population size to other control parameters of the genetic algorithm.

In section 2 , we introduce our notation and propose the sampling error criteria for sizing the population. Section 3 studies the GA's sampling behavior over a general class of subsets of the individual space and applies the result to a special class of subsets—schemata to establish lower bounds on the population size. We conclude in Section 4 with some discussions.

2 GENETIC ALGORITHMS, SAMPLING DISTRIBUTIONS AND SAMPLING ERRORS

We consider GAs with a binary string representation. Let N be the population size and l be the encoding length. Each individual in a population corresponds to an element of the *individual space* $S = \{0,1\}^l$. The *population space* is denoted by $S^N = S \times S \times \cdots \times S$

and a population $\mathcal{X} \in S^N$ is understood as a set of individuals $\mathcal{X} = (X_1, X_2, ..., X_N)$.

A schema $\mathcal{L}$ is a hyperspace in the individual space S. The *order* $o(\mathcal{L})$ and the *defining length* $\delta(\mathcal{L})$ of a schema $\mathcal{L}$ are defined respectively to be the number of fixed positions in the schema and the distance between the first and last defining positions.

To simplify the discussion, we consider the canonical genetic algorithm (CGA) that uses the standard proportional selection, one-point crossover, and bit-wise mutation operators and has a fixed population size N. Let $\{\mathcal{X}(k), k \geq 0\}$ be the sequence of populations in CGA. Given the current population $\mathcal{X}(k)$, CGA uses these genetic operators to generate the next population $\mathcal{X}(k+1)$, one individual at a time. An individual is generated by first selecting two individuals from the current population, applying the crossover operator to the two selected individuals, and then randomly picking one of the two resulting individuals of the crossover to apply the mutation operator. The above operation can be abstracted as an operator T that is basically a (random) transformation from the population space into the individual space. Given the current population $\mathcal{X}$, the distribution of $T(\mathcal{X})$ is called the conditional sampling distribution of CGA

$$p(\mathcal{X}, Y) \triangleq P\{T(\mathcal{X}) = Y\}, \quad Y \in S. \tag{1}$$

Let $f : S \to R^+$ be the fitness function, p_c be the crossover probability, and p_m the mutation probability. For any subset $\mathcal{C}$ in the individual space S, let $N(\mathcal{X}(k), \mathcal{C})$ be the number of individuals in $\mathcal{C}$ and

$$p(\mathcal{X}(k), \mathcal{C}) \triangleq \sum_{Y \in \mathcal{C}} P(\mathcal{X}(k), Y) \tag{2}$$

be the sampling probability of $\mathcal{C}$ in the next generation. Holland's schema theorem states that small, low-order schemata with above-average performance are allocated exponentially increasing trials in the next generation [Holland, 1975]. That is, for any schema $\mathcal{L}$,

$$E[N(\mathcal{X}(k+1), \mathcal{L})] \geq N(\mathcal{X}(k), \mathcal{L}) \cdot \frac{\overline{f}(\mathcal{X}(k), \mathcal{L})}{\overline{f}(\mathcal{X}(k))}(1 - p_c \frac{\delta(\mathcal{L})}{l-1} - o(\mathcal{L})p_m) \tag{3}$$

where $E[\cdot]$ denotes the mathematical expectation, $\overline{f}(\mathcal{X}(k))$ is the average fitness of $\mathcal{X}(k)$, and $\overline{f}(\mathcal{X}(k), \mathcal{L})$ is the average fitness of the individuals of $\mathcal{L}$ in $\mathcal{X}(k)$.

Since given the current population $\mathcal{X}(k)$, the individuals in the next generation are conditionally independent and have an identical distribution $p(\mathcal{X}(k), \cdot)$, we see that $N(\mathcal{X}(k+1), \mathcal{L})$ has the binomial distribution with the parameters N and

$$p(\mathcal{X}(k), \mathcal{L}) = \sum_{Y \in \mathcal{L}} P(\mathcal{X}(k), Y). \tag{4}$$

The sampling distribution $p(\mathcal{X}(k), \cdot)$ comprises all the relevant information about the current environment (i.e., the structure and the fitness information of the current population) that can be used by CGA to guide future search. Under the belief that the mechanism of natural evolution that the genetic algorithms try to mimic is a good one, we should sample the regions of the individual space according to the proportion suggested by $p(\mathcal{X}(k), \cdot)$. It is not the case if the population size N is finite. Because of sampling error (or noise), the relative frequency $\frac{N(\mathcal{X}(k+1), \mathcal{L})}{N}$ of the samples in the next generation may deviate considerably from the probability $p(\mathcal{X}(k), \mathcal{L})$.

Based on this observation and motivated from the study of computational learning theory, we propose the following two types of sampling error criteria for bounding the population size. The basic idea underlying our criteria is to find effective population sizes so that, with probability close to 1, the deviation (or the relative deviation) of the relative sample frequencies from the true sampling probabilities is small uniformly over a class of schemata.

Definition 2.1 *Let ϵ and δ be small positive scalars, called respectively the accuracy parameter and the confidence parameter. A population size N is said to be (ϵ, δ, m)-effective if for any $k \geq 1$ and $\mathcal{X}(k) \in S^N$, we have*

$$P\{\sup_{1 \leq o(\mathcal{L}) \leq m} |\frac{1}{N} N(\mathcal{X}(k+1), \mathcal{L}) - p(\mathcal{X}(k), \mathcal{L})| \geq \epsilon\} < \delta, \tag{5}$$

where the sup is taken over all the schemata $\mathcal{L}$ with the order $1 \leq o(\mathcal{L}) \leq m$.

There are however some problems with the absolute sample error criterion in Definition 2.1. For example, let $\mathcal{X}$ be a population and $\mathcal{L}_j^1$ be a schema of order one with the defining gene position j and the defining allele 1. Suppose $p(\mathcal{X}, \mathcal{L}_j^1)$ is very small so that, with a relatively large probability, $\frac{1}{N} N(\mathcal{X}(k+1), \mathcal{L}_j^1) = 0$, i.e., the allele 1 at gene position j is lost in the next generation. Since allele loss is related to the problem of premature convergence ([Leung *et al.*, 1997]), a population size satisfying (5) should not be viewed as a good choice. However, according to Definition 2.1, such a population size may be (ϵ, δ, l)-effective since

$$\left| \frac{1}{N} N(\mathcal{X}(k+1), \mathcal{L}_j^1) - p(\mathcal{X}(k), \mathcal{L}_j^1) \right|$$

$$= \left|0 - p(\mathcal{X}, \mathcal{L}_j^1)\right| \leq \epsilon.$$

To overcome this kind of problems, we can consider the following criterion.

Definition 2.2 *Let ϵ and δ be small positive scalars, called respectively the accuracy parameter and the confidence parameter. A population size N is said to be strongly (ϵ, δ, m)-effective if for any $k \geq 1$ and $\mathcal{X}(k) \in S^N$, we have*

$$P\{\sup_{1 \leq o(\mathcal{L}) \leq m} \frac{1}{\sqrt{p(\mathcal{X}(k), \mathcal{L})}} \left|\frac{1}{N} N(\mathcal{X}(k+1), \mathcal{L}) - p(\mathcal{X}(k), \mathcal{L})\right| \geq \epsilon\} < \delta, \quad (6)$$

where the sup is taken over all the schemata $\mathcal{L}$ with the order $1 \leq o(\mathcal{L}) \leq m$.

3 LOWER BOUNDS ON THE POPULATION SIZE

In this section, we establish lower bounds on the effective population size by using tools from the theory of uniform strong law of large numbers [Blumer *et al.*, 1989, Haussler, 1992, Pollard, 1984]. We first present a more general theorem on the deviation of the relative sampling frequencies from the true sampling probabilities.

Theorem 3.1 *Let $\mathbf{C}$ be a class of subsets in the individual space with the cardinality $|\mathbf{C}|$. For any $0 < \epsilon, \delta < 1$, if the population size N satisfies*

$$N \geq \frac{2}{\epsilon^2}(\ln|\mathbf{C}| + \ln\frac{2}{\delta}), \quad (7)$$

we have

$$P\left\{\sup_{\mathcal{C} \in \mathbf{C}} \left|\frac{1}{N} N(\mathcal{X}(k+1), \mathcal{C}) - p(\mathcal{X}(k), \mathcal{C})\right| \geq \epsilon\right\} < \delta, \quad (8)$$

Proof: A standard application of the uniform strong law of large numbers based on Hoeffding's inequality [Pollard, 1984]. See [Gao, 2002] for the details.

Taking $\mathbf{C}$ to be the class of schemata, we obtain the following corollary which gives a lower bound on (ϵ, δ)-effective population size.

Corollary 3.1 *Let $0 < \epsilon, \delta < 1$ be the given accuracy and confidence parameters, and l be the encoding length (problem size). Then, any population size satisfying*

$$N \geq N(\epsilon, \delta, l) \triangleq \frac{2}{\epsilon^2}(l \cdot \ln 3 + \ln\frac{2}{\delta}), \quad (9)$$

is (ϵ, δ, l)-effective.

Since for fixed (ϵ, δ), the lower bound $N(\epsilon, \delta, l)$ is a linear in the problem size l, We have the following corollary which depicts GAs' implicit parallelism from a new perspective.

Corollary 3.2 *To process all the schemata whose number increases exponentially with the problem size l, the effective population size N only needs to increase linearly with the problem size.*

In the following, we establish a lower bound under criterion proposed in Definition 2.2.

Definition 3.1 *Let $f : S \to R^+$ be the fitness function. Given a population $\mathcal{X} = (X_1, \cdots, X_N)$, $X_i = (x_{i1}, \cdots, x_{il})^T$, $1 \leq i \leq N$, for any positive integer $1 \leq j \leq l$, let I_j^1 and I_j^0 be the sets of indices of the individuals in $\mathcal{X}$ that have one or zero at the jth component (gene position) respectively. Write $F(\mathcal{X}) = \sum_{i=1}^N f(X_i)$ and $F_j^1(\mathcal{X}) = \sum_{i \in I_j^1} f(X_i)$. We call*

$$a_j(\mathcal{X}) = \frac{F_j^1(\mathcal{X})}{F(\mathcal{X})}, \quad \text{and} \quad b_j(\mathcal{X}) = \frac{F_j^0(\mathcal{X})}{F(\mathcal{X})}$$

the relative fitness of the order one schema $\mathcal{L}_j^1$ and $\mathcal{L}_j^0$ respectively.

Theorem 3.2 *Let $0 < \epsilon, \delta < 1$ be the given accuracy and confidence parameters, l be the encoding length (problem size), p_m be the mutation probability, and $f : S \to R^+$ the fitness function. For a given population $\mathcal{X}$, let $\overline{p} = \min_{1 \leq j \leq l} \left(\frac{1}{2} - 2|a_j(\mathcal{X}) - \frac{1}{2}||p_m - \frac{1}{2}|\right) > 0$. Then, any population size satisfying*

$$N \geq N(\epsilon, \delta, l, \mathcal{X}, f) \triangleq \frac{1 - \overline{p} + \frac{1}{3}\epsilon}{2\overline{p}} \cdot \frac{1}{\epsilon^2}(\ln l + \ln\frac{4}{\delta}), \quad (10)$$

is strongly $(\epsilon, \delta, 1)$-effective.

Proof: The proof is based on Bernstein's inequality [Pollard, 1984] and an explicit formula for the conditional probability of order 1 schema $p(\mathcal{X}, \mathcal{L}_j^1) = a_j(\mathcal{X}) + (1 - 2a_j(\mathcal{X}))p_m$ which can be established similar to the results in [Leung *et al.*, 1997]. See [Gao, 2002] for the details. It can be seen the lower bound $N(\epsilon, \delta, l, \mathcal{X}, f)$ increases in $|a_j^1 - \frac{1}{2}|$ and $|p_m - \frac{1}{2}|$. This is in consistence with the experimental observation in [Leung *et al.*, 1997, Schaffer *et al.*, 1989].

4 CONCLUDING REMARKS

The theory of uniform strong law of large numbers has found many applications in the fields of statistics and

machine learning. It is the basis of the Probably Approximately Correct (PAC) model of machine learning ([Blumer *et al.*, 1989, Pollard, 1984]). The work presented in this paper shows that these techniques can also be applied to the analysis of genetic algorithms. This is perhaps not a surprise since the process of evolution is in essence a learning process.

It is well known that the population size is one of the most important control parameters in the design of genetic algorithms [Eiben *et al.*, 1999]. In recent years, much attention has been paid to the design of sampling-based genetic algorithms that generate new solutions by sampling the probability distribution estimated from the current population [Leung *et al.*, 2001, Muhlenbein and Mahnig, 1999, Pelikan *et al.*, 1999]. From a statistics and machine learning point of view, an appropriate population size is even more critical to the success of this type of algorithms. In this paper, we have proposed two sampling error criteria and established corresponding lower bounds. We hope studies along this direction will help us gain deeper understanding about the nature and limitation of the sampling-based genetic algorithms.

References

[Bertoni and Dorigo, 1993] A. Bertoni and M. Dorigo. Implicit parallelism in genetic algorithms. *Artificial Intelligence*, 61:307–314, 1993.

[Blumer *et al.*, 1989] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

[Eiben *et al.*, 1999] A.E. Eiben, R. Hinterding, and Z. Michalewicz. Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2):124–141, 1999.

[Gao, 2002] Y. Gao. Lower bounds on the population size in genetic algorithms and implicit parallelism revisited. Technical Report TR02-20, Department of Computing Science, University of Alberta, 2002.

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

[Grefenstette, 1986] J. J. Grefenstette. Optimization of control parameters for genetic algorithms. *IEEE Trans. on Systems, Man, and Cybernetics*, 16(1):122–128, 1986.

[Harik *et al.*, 1997] G. Harik, E. Cantu-Paz, D. E. Goldberg, and B. L. Miller. The gambler's ruin problem, genetic algorithms, and the sizing of populations. In *Proc. of the 1997 IEEE International Conf. on Evol. Comput.*, pages 7–12. IEEE Press, 1997.

[Haussler, 1992] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[Holland, 1975] J. H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, 1975.

[Leung *et al.*, 1997] Y. Leung, Y.Gao, and Z. Xu. Degree of population diversity: A perspective on premature convergence in genetic algorithms and its markov chain analysis. *IEEE Trans. on Neural Networks*, 8(5):1165–1176, 1997.

[Leung *et al.*, 2001] Y. Leung, Y. Gao, and W. Zhang. A genetic-based method for training fuzzy systems. In *Proc. of the 10th IEEE International Conference on Fuzzy Systems*, volume 1, pages 123–126. IEEE, 2001.

[Muhlenbein and Mahnig, 1999] H. Muhlenbein and Th. Mahnig. Convergence theory and applications of the factorized distribution algorithm. *Journal of Computing and Information Technology*, 7:19–32, 1999.

[Pelikan *et al.*, 1999] M. Pelikan, D. E. Goldberg, and F. Lobo. A survey of optimization by building and using probabilistic models. Technical Report 99018, IlliGAL, Univ. of Illinois, 1999.

[Pollard, 1984] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[Schaffer *et al.*, 1989] J. D. Schaffer, R. A. Caruana, L. J. Eshelman, and R. Das. A study of control parameters affecting online performance of genetic algorithms for function optimization. In J. D. Schaffer, editor, *Proc. of the Third International Conference on Genetic Algorithms*, pages 51–60. Morgan Kaufmann, 1989.