Lower Bounds on the Population Size in Genetic Algorithms and Implicit Parallelism Revisited

Yong Gao
Department of Computing Science
University of Alberta
Edmonton, AB, Canada T6G 2H1 and
e-mail: ygao@cs.ualberta.ca

ABSTRACT

Determining an appropriate population size is very important in genetic algorithms and is closely related to the principle of implicit parallelism. In this report, the problem of bounding the population size is formulated as that of minimization of sampling errors. Two sampling error criteria are proposed for bounding the population size in genetic algorithms. A theorem on the sampling error of genetic algorithms over a general class of subsets in the individual space is established. Applying the result to the class of schemata, we derive two kinds of lower bounds on the population size and present the principle of implicit parallelism from a new perspective. It is further shown that the lower bound can also result in the monotonic convergence of the correct schema. The lower bounds also depict how the necessary population size is related to the mutation probability and some population statistics. In particular, our results give an explanation to an experimental observation of Schaffer et al [21].

Keywords: Genetic algorithms, Population size, Uniform strong law of large numbers, Sampling error, Implicit parallelism

I. Introduction

Genetic algorithms (GAs) are robust and efficient search techniques inspired by Darwin's theory of natural evolution [1], and have been applied with success to a wide variety of function optimization, machine learning, and engineering control problems [1]-[5]. Apart from the large volumes of experimental work in the literature, much progress has been made recently in the theoretical investigation of GAs [1], [5]-[20]. Among the various analytic approaches, schema analysis, introduced first by Holland [1] and further developed by the others [11]-[16], is still one of the most powerful methods to describe the actual behavior of GAs.

From his schema theorem, Holland derived the famous principle of implicit parallelism by counting the number of schemata that can be effectively processed. Holland's result on implicit parallelism is usually quoted in such a way that the number of schemata processed effectively is proportional to the cube of the population size N, i.e., $O(N^3)$. This result, although it sounds attractive, is misleading. In fact, there is a gap in Holland's original derivation, as has been noted by Goldberg [17] and Bertoni and Dorigo [16]. A closer look at the original derivation shows that the $O(N^3)$ estimation is valid only when the population size N and the encoding length l have the relation $N = c \cdot 2^l$. Bertoni and Dorigo in [16] had further demonstrated that different orders of estimations could be obtained under different assumptions on the relations between N and l. Since implicit parallelism is widely recognized as the most significant feature of GAs, the result of Bertoni and Dorigo [16] raises the following serious question on the reasonableness of the current Holland-type interpretation of GAs' implicit parallelism: does there exist a universal order of estimation on implicit parallelism that is valid uniformly for all population sizes?

Another important problem closely related to implicit parallelism is that of bounding the population size. There is significant empirical evidence showing that the larger the population size, the better the ultimate convergence quality of GAs [8],[17],[21],[22]. On the other hand, it is well-accepted that a large population will result in both slow convergence and extra computational effort [17], [22], [23]. There has been some theoretical work dealing with the problem of bounding the population size. To determine the appropriate population size, the process of selection is usually viewed as a game played among a class of competing schemata that forms a partition of the individual space [1], [14], [17]-[20]. Goldberg et al [18],[19], derived population-sizing equations by controlling the error probability on a single trial to be under a pre-specified level. There are, however, some problems with Goldberg's population-sizing equations.

(1) The equations are derived for a pre-specified class of competing schemata, and different classes of competing schemata will result in different equations. The problem is which class of competing schemata one should choose when determining the population size.

(2) The equations require some knowledge about the tasks to be solved. Goldberg et al [18], [19], suggest using the online population measurements to "estimate" the required knowledge. This approach, however, leaves us in an embarrassing situation, since from the statistics point of view, we have to bound the population size for the second time to make sure that these estimates are good approximations to the required knowledge.

(3) The equations only consider the first generation of the run under the belief that if decisions are correct in the first generation, then GAs will converge to the right building

blocks.

The last problem was recently tackled by Harik, et al. [20], where the convergence quality of GAs (in fact, the convergence of a particular class of competing schema) is also taken into account when establishing the population size equations. Under the assumption of mutual independence between schemata, a population size is derived from an expression of absorbing probability of the correct schemata.

Besides minimizing the sampling errors, the interaction among the control parameters of GAs must also be considered when determining the population size. Schaffer et al [21] systematically studied the combined influences of control parameters on the on-line performance of GAs by a series of elaborate experiments. An important observation made in their experiments is that the population size and the mutation probability are in inverse relation for the optimal on-line performance on the class of test functions they used. To the best of our knowledge, there is so far no theoretical result to explain this observation.

In this paper, we will deal with the above mentioned problems of implicit parallelism and the population size in a unified approach. Using tools from the theory of uniform strong law of large numbers for random variables, we study the sampling behavior of GAs in details and establish lower bounds on the population size under certain sampling error criteria. Our first lower bound given under the criterion that minimizes the absolute sampling error uniformly over all schemata, is a linear function of the problem size (encoding length) l. The lower bound is independent of the fitness function and the current state of the population, and hence is universal. Based on the first lower bound, we can interpret the implicit parallelism of GAs from a new perspective, i.e.,

> *To effectively process all the schemata whose number increases exponentially with the problem size, the population size for GAs only needs to increase linearly with the problem size.*

This interpretation avoids the inconsistency in Holland-type interpretations of implicit parallelism of GAs.

We obtain our second lower bound under a criterion that minimizes the relative sampling error uniformly over the schemata of order one. The second lower bound depends on the mutation probability and some population statistics. This is contrary to Goldberg's population-sizing equations that depend on the mean fitness and the fitness variances of the schemata, which can only be approximated by the population statistics. Our second lower bound also explains theoretically the experimental observation of Schaffer et al [21] that the population size and the mutation probability have an inverse relation for the optimal online performance of GAs.

By virtue of the uniform strong law of large numbers, the lower bound result is universal in that it does not depend on particular population states and is valid for all schema. Hence, for GAs with the population size larger than the lower bound, sampling errors are controlled at a specific level throughout the run of GAs. We shall use this observation, in conjunction with schema theorem, to establish some results on the convergence of GAs.

The remainder of the paper is organized as follows. In section II, we first introduce some notation and concepts, and then propose the sampling error criteria for sizing the population. Section III studies the GA's sampling behavior over a general class of subsets of the individual space and uses the result to a special class of subsets—schemata to establish lower bounds on the population size. Under certain assumptions on the initial population and the accuracy parameters, a convergence result for the fittest schema is also derived. The lower bound results are then used to present our new interpretation of implicit parallelism, and to explain the empirical observation made by Schaffer et al [21]. In section V, we conclude and present some directions for future work.

II. Genetic Algorithms, Sampling Distributions and Sampling Error Criteria for the Population Size

We consider GAs with binary string representations. Let N be the population size and l be the encoding length (problem size). Each individual in a population corresponds to an element of the space $S = \{0,1\}^l$, which is called the *individual space*. Denote the *population space* by $S^N = S \times S \times \cdots \times S$. For the sake of convenience, we write a population $\vec{X} \in S^N$ in both the vector and matrix forms as follows:

$$\vec{X} = (X_1, X_2, ..., X_N)^T = \begin{pmatrix} x_{11}\ x_{12}...x_{1l} \\ x_{21}\ x_{22}...x_{2l} \\ .... \\ x_{N1}\ x_{N2}...x_{Nl} \end{pmatrix}$$

where $X_i \in S$ is the ith individual of $\vec{X}$, while x_{ij} is the jth component of X_i.

A schema [1] $\mathcal{L}$ 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 [2]. For a population $\vec{X} \in S^N$, let $N(\vec{X}, \mathcal{L})$ denote the number of the individuals of $\vec{X}$ that are contained in $\mathcal{L}$.

We consider the following abstract genetic algorithm. Following Leung, Gao, and Xu [8], we call it the *Canonical genetic algorithm (CGA)*.

CGA

Step 1: Set $k = 0$ and generate initial population $\vec{X}\,(0)$;

Step 2: Independently select N pairs of individuals from the current population for reproduction;

Step 3: Independently perform crossover on the N selected pairs of individuals to generate N new intermediate individuals;

Step 4: Independently mutate the N intermediate individuals to get the next generation $\vec{X}\,(k+1) = (X_1(k+1), \cdots, X_N(k+1))$;

Step 5: Stop if some stopping criterion is met. Else, set $k = k+1$ and go to Step 2.

From the mathematical point of view, genetic operators are random mappings between the spaces S^N, S^2, and S. They are the analogous idealized abstractions of some of the genetic mechanisms in the evolution of natural organisms. To facilitate our later analysis, we present in the following the probabilistic definitions of several basic operators.

(a) The proportional selection operator, $T_s : S^N \longrightarrow S^2$, selects a couple of parents from the given population for reproduction. Given the population $\vec{X}$, the probability of selecting $(X_i, X_j) \in S^2$ as the parents is

$$P\{T_s(\vec{X}) = (X_i, X_j)\} = \frac{f(X_i)}{\sum_{X \in \vec{X}} f(X)} \cdot \frac{f(X_j)}{\sum_{X \in \vec{X}} f(X)}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq N. \qquad (1)$$

(b) The crossover operator, $T_c : S^2 \longrightarrow S$ generates an individual from the selected parents. Given the parent $X_i = (x_{i1}, \cdots, x_{il})$, $i = 1, 2$, the probability for the one-point crossover operator to generate an individual Y is

$$P\{T_c((X_1, X_2)) = Y\} = \begin{cases} \frac{k \cdot p_c}{l}, & \text{if } Y \neq X_1 \\ (1 - p_c) + \frac{k \cdot p_c}{l}, & \text{if } Y = X_1 \end{cases} . \qquad (2)$$

where $0 \leq p_c \leq 1$ is the so-called crossover probability, k is the number of crossover points at which the crossover of X_1 and X_2 can generate Y.

Remark 2.1: Throughout the present paper, we only consider the 2-parents-1-child crossover operator, i.e., crossing two parents will always generate only one child (which we assume to be the first offspring). Such a scheme has also been adopted by many other theoretical works on GAs [1]. This is because only with 2-parents-1-child crossover, the offspring in the next generation are conditionally independent and identically distributed. This is crucial in the derivation of our results from the strong law of large numbers.

Remark 2.2: Though only the one-point crossover operator is defined in the above, all the results in this paper are valid for any kind of crossover operators such as multi-point crossover, uniform crossover, etc.

(c) The mutation operator, $T_m : S \longrightarrow S$, operates on the individual by independently perturbing each bit string in a probabilistic manner and can be specified as follows:

$$P\{T_m(X) = Y\} = p_m^{|X-Y|}(1 - p_m)^{l-|X-Y|}. \qquad (3)$$

Definition 2.1 Let $\{\vec{X}(k), k \geq 1\}$ be the sequence of populations of GAs. Denote by

$$P(\vec{X}, \vec{Y}) \triangleq P\{\vec{X}(k+1) = \vec{Y} \mid \vec{X}(k) = \vec{X}\}$$

the conditional probability of the event that the population of the next generation is $\vec{Y}$ given the current population $\vec{X}(k) = \vec{X}$. The (conditional) sampling distribution of the CGA is defined to be

$$p(\vec{X}, Y) \triangleq P\{T_m(T_c(T_s(\vec{X}(k)))) = Y | \vec{X}(k) = \vec{X}\}, \quad Y \in S.$$

Remark 2.3: From the description of CGA, it is clear that the individuals $X_i(k+1)$, $1 \leq k \leq N$, are conditionally independent and identically distributed given the current population $\vec{X}(k) = \vec{X}$. Hence, we have for each $1 \leq i \leq N$,

$$P\{X_i(k+1) = Y | \vec{X}(k) = \vec{X}\} = p(\vec{X}, Y).$$

That is, $p(\vec{X}, Y)$ is the common conditional distribution of the individuals in generation $k+1$.

We now begin to derive our sampling error criteria for the population size. To get some motivations, let us first recall Holland's schema theorem. Let $f : S \to R^+$ be the fitness function, p_c be the crossover probability, and p_m be the mutation probability. For any subset $\mathcal{C}$ in the individual space S, let $N(\vec{X}(k), \mathcal{C})$ be the number of individuals in $\mathcal{C}$ and

$$p(\vec{X}(k), \mathcal{C}) \triangleq \sum_{Y \in \mathcal{C}} P(\vec{X}(k), Y) \tag{4}$$

be the sampling probability of $\mathcal{C}$ in the next generation. The schema theorem states that small, low-order schemata with above-average performance are allocated exponentially increasing trials in the next generation [2]. That is, for any schema $\mathcal{L}$, we have

$$E[N(\vec{X}(k+1), \mathcal{L})] \geq N(\vec{X}(k), \mathcal{L}) \cdot \frac{\overline{f}(\vec{X}(k), \mathcal{L})}{\overline{f}(\vec{X}(k))}(1 - p_c\frac{\delta(\mathcal{L})}{l-1} - o(\mathcal{L})p_m) \tag{5}$$

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

Consider the random variable $N(\vec{X}\ (k+1), \mathcal{L})$. Since given the current population $\vec{X}\ (k)$, the individuals, $X_i(k+1), \quad 1 \leq i \leq N$, are conditionally independent and have a common distribution $p(\vec{X}\ (k), \cdot)$, we see that $N(\vec{X}\ (k+1), \mathcal{L})$ has the binomial distribution with the parameters N and

$$p(\vec{X}\ (k), \mathcal{L}) = \sum_{Y \in \mathcal{L}} P(\vec{X}\ (k), Y). \tag{6}$$

The sampling distribution $p(\vec{X}\ (k), \cdot)$ 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 the CGA to guide the future search. Ideally, the algorithm should sample the regions of the individual space according to the proportion suggested by $p(\vec{X}\ (k), \cdot)$. Unfortunately, it is not the case if the population size N is finite. Because of sampling error (or noise), the relative frequency $\frac{N(\vec{X}(k+1), \mathcal{L})}{N}$ of the samples in the next generation may deviate considerably from the probability $p(\vec{X}\ (k), \mathcal{L})$. This deviation can cause the performance of the algorithm to deteriorate.

By the classical strong law of large numbers, for any $\epsilon, \delta > 0$, there exists an $N(\epsilon, \delta, \mathcal{L}, \vec{X}\ (k))$ such that for all $N \geq N(\epsilon, \delta, \mathcal{L}, \vec{X}\ (k))$,

$$P\{|\frac{1}{N} N(\vec{X}\ (k+1), \mathcal{L}) - p(\vec{X}\ (k), \mathcal{L})| > \epsilon\} < \delta. \tag{7}$$

This means that for a particular schema and a population $\vec{X}\ (k)$, the probability that the relative frequency has a large deviation from $p(\vec{X}\ (k), \mathcal{L})$ may be arbitrarily small if the population size is sufficiently large. We are, however, not just content with this. Because we do not know a priori the global optimal solution and because the sample distribution $p(\vec{X}(k), \cdot)$ depends on the structure of the population $\vec{X}(k)$, we would like (7) to be valid uniformly for all schemata (or at least a class of schemata) and all population states.

Based on these considerations, 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 2.2 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 weakly (ϵ, δ, m)-effective if for any $k \geq 1$ and $\vec{X}(k) \in S^N$, we have

$$P\left\{\sup_{1\leq o(\mathcal{L})\leq m}\left|\frac{1}{N}N(\vec{X}(k+1), \mathcal{L}) - p(\vec{X}(k), \mathcal{L})\right| \geq \epsilon\right\} < \delta, \tag{8}$$

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

Definition 2.3 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 $\vec{X}(k) \in S^N$, we have

$$P\left\{\sup_{1\leq o(\mathcal{L})\leq m}\frac{1}{\sqrt{p(\vec{X}(k), \mathcal{L})}}\left|\frac{1}{N}N(\vec{X}(k+1), \mathcal{L}) - p(\vec{X}(k), \mathcal{L})\right| \geq \epsilon\right\} < \delta, \tag{9}$$

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

III. Lower Bounds on the Population Size and a New Perspective of Implicit Parallelism

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 [24],[25]. 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

$$N \geq \frac{2}{\epsilon^2}(\ln |\mathbf{C}| + \ln \frac{2}{\delta}), \tag{10}$$

then we have

$$P \left\{ \sup_{\mathcal{C} \in \mathbf{C}} \left| \frac{1}{N} N(\vec{X}\ (k+1), \mathcal{C}) - p(\vec{X}\ (k), \mathcal{C}) \right| \geq \epsilon \right\} < \delta, \tag{11}$$

Proof. For a fixed $\mathcal{C} \in \mathbf{C}$, let

$$\xi_i = I_{\mathcal{C}}(X_i(k+1)), \quad 1 \leq i \leq N,$$

where $I_{\mathcal{C}}(\cdot)$ is the indicator function of $\mathcal{C}$. Since $X_i(k+1), \quad 1 \leq i \leq N$, are conditionally independent (given $\vec{X}\ (k)$), and have the common distribution $p(\vec{X}\ (k), \cdot)$, we see that $(\xi_i, 1 \leq i \leq N)$ are conditionally independent Bernoulli random variables with the common success probability $p(\vec{X}\ (k), \mathcal{C})$. It follows that for each $1 \leq i \leq N$,

$$0 \leq \xi_i \leq 1, \text{and} \ \ 0 \leq E[\xi_i] = p(\vec{X}\ (k), \mathcal{C}) \leq 1.$$

Recall that $N(\vec{X}\ (k+1), \mathcal{C})$ is the number of the individuals of $\vec{X}\ (k+1)$ contained in $\mathcal{C}$, so

$$\frac{1}{N} N(\vec{X}\ (k+1), \mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} \xi_i.$$

By Hoeffding's inequality (See Lemma A.1 in Appendix A), we obtain

$$\begin{aligned}
& P\left\{\left|\frac{1}{N} N(\vec{X}\ (k+1), \mathcal{C})-p(\vec{X}\ (k), \mathcal{C})\right|>\epsilon\right\} \\
& =P\left\{\left|\frac{1}{N} \sum_{i=1}^{N} \xi_i-p(\vec{X}\ (k), \mathcal{C})\right|>\epsilon\right\} \\
& =P\left\{\left|\sum_{i=1}^{N}(\xi_i-p(\vec{X}\ (k), \mathcal{C}))\right|>N \epsilon\right\} \\
& \leq 2 \exp \left[-\frac{N \epsilon^2}{2}\right].
\end{aligned}$$

Note that the right hand side of the above inequality is independent of $\mathcal{C}$. Now, suppose that N is a population size with

$$N \geq N(\epsilon, \delta, l) \triangleq \frac{2}{\epsilon^2}(\ln|\mathcal{C}|+\ln \frac{2}{\delta}).$$

we get

$$\begin{aligned}
& P\left\{\sup_{\mathcal{C} \in \mathbf{C}}\left|\frac{1}{N} N(\vec{X}\ (k+1), \mathcal{C})-p(\vec{X}\ (k), \mathcal{C})\right| \geq \epsilon\right\} \\
& \leq \sum_{\mathcal{C} \in \mathbf{C}} 2 \exp(-\frac{N \epsilon^2}{2}) \\
& =2 \cdot|\mathbf{C}| \exp \left[-\frac{N \epsilon^2}{2}\right] \\
& \leq \delta.
\end{aligned}$$

This completes the proof. Q.E.D.

Taking $\mathbf{C}$ to be the class of schemata, we obtain the following corollary which gives a lower bound on the weakly (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

$$N \geq N(\epsilon, \delta, l) \triangleq \frac{2}{\epsilon^2}(l \cdot \ln 3+\ln \frac{2}{\delta}), \tag{12}$$

is weakly (ϵ, δ, l)-effective.

Proof. If $\mathbf{C}$ is the set of all the schemata in the individual space, then its cardinality is 3^l. The corollary follows. Q.E.D.

From corollary 3.1, we can see that the lower bound $N(\epsilon, \delta, l)$ on the effective population size depends only on the problem size l and the pre-specified accuracy and confidence parameters (ϵ, δ). Thus the lower bound is valid for any problem to be solved and any population state, and hence is universal. Moreover, it can be observed that for fixed (ϵ, δ), the lower bound $N(\epsilon, \delta, l)$ is a linear function of the problem size l. On the other hand, simple calculation shows that there are 3^l schemata in the individual space with the encoding length l. We therefore obtain 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. That is, $N = O(l)$.

The universal lower bound presented in Corollary 3.1 is, however, too conservative to be of practical use, particularly when the accuracy parameter ϵ is very small. In fact, $N(\epsilon, \delta, l)$ is a worst-case lower bound in the sense that the criterion (8) must be fulfilled for all schemata. In practice, we may only be concerned with a special class of schemata such as the schemata of order 1. In this case, an application of theorem 3.1 yields a lower bound for the population size as

$$N(\epsilon, \delta, 1) = \frac{2}{\epsilon^2}(\ln l + \ln \frac{2}{\delta}),$$

which would be more reasonable. Moreover, as will be shown in corollary 3.3, the accuracy parameter ϵ need not be extremely small to obtain some kinds of monotonic convergence results of GAs if the signal difference between competing schemata is significant.

Another advantage of the universality of our lower bound is that it can assure the ultimate convergence of the correct schema. In the following we will show that under certain assumptions on the initial population and the accuracy parameter ϵ, the number of individuals in the fittest schema will monotonously increase during the run of a GA with probability close to 1. To facilitate the presentation, let us rewrite the schema theorem as

$$E[N(\vec{X}\ (k+1), \mathcal{L})] \geq N(\vec{X}\ (k), \mathcal{L}) \cdot a(\mathcal{L}) \cdot d \tag{13}$$

where $a(\mathcal{L})$ is the fitness ratio of the schema $\mathcal{L}$ relative to the average fitness and d is the disruptive effect of crossover and mutation.

Corollary 3.3 Consider the schema $\mathcal{L}$ of order m with the highest fitness ratio among the same class of competing schemata(i.e., those schemata with the same order and defining positions). Suppose that the initial population is uniformly distributed and the accuracy parameter ϵ satisfies

$$\epsilon < \frac{1}{2^m}(a(\mathcal{L})d - 1). \tag{14}$$

If the population size is weakly (ϵ, δ, l)-effective, then with probability $1 - \delta$, we have

$$N(\vec{X}\ (k+1), \mathcal{L}) \geq N(\vec{X}\ (k), \mathcal{L}), \quad \forall k \geq 1 \tag{15}$$

provided that $N(\vec{X}\ (k), \mathcal{L}) < N$.

Proof. If the population size N is (ϵ, δ, l)-effective, then with probability $1 - \delta$,

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

for all $k \geq 1$. It follows that

$$\left| N(\vec{X}\ (k+1), \mathcal{L}) - E[N(\vec{X}\ (k+1), \mathcal{L})] \right| < \epsilon N.$$

So according to (13) and the assumption (14), we have

$$\begin{aligned} & N(\vec{X}\ (k+1) \geq N(\vec{X}\ (k), \mathcal{L}) \cdot a(\mathcal{L}) \cdot d - \epsilon N \\ & = N(\vec{X}\ (k), \mathcal{L}) + N(\vec{X}\ (k), \mathcal{L})(a(\mathcal{L}) \cdot d - 1) - \epsilon N \\ & \geq N(\vec{X}\ (k), \mathcal{L}) + N(\vec{X}\ (k), \mathcal{L})(a(\mathcal{L}) \cdot d - 1) - \frac{1}{2^m} N(a(\mathcal{L}) \cdot d - 1). \qquad (16) \end{aligned}$$

We now show that

$$N(\vec{X}\ (k), \mathcal{L}) \geq \frac{1}{2^m} N, \forall k \geq 1. \qquad (17)$$

By assumption, for the initial population, we have $N(\vec{X}\ (1), \mathcal{L}) \geq \frac{1}{2^m} N$. Suppose that $N(\vec{X}\ (k), \mathcal{L}) \geq \frac{1}{2^m} N$, then by (16),

$$N(\vec{X}\ (k+1), \mathcal{L}) \geq N(\vec{X}\ (k), \mathcal{L}) \geq \frac{1}{2^m} N$$

By method of induction, (17) follows. This, together with (16), proves the corollary. Q.E.D.

There are, however, some problems with the absolute sample error criteria in definition 2.2. For example, let $\vec{X}$ 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 $p(\vec{X}, \mathcal{L}_j^1)$ is very small so that, with a relatively large probability,

$$\frac{1}{N} N(\vec{X}\ (k+1), \mathcal{L}_j^1) = 0, \qquad (18)$$

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 ([8]), a population size satisfying (18) should not be viewed as a good choice. However, according to the absolute sample error criteria, such a population size may be weakly (ϵ, δ, l)-effective since

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

To overcome the above problem with our lower bound on the population size, we will now establish another lower bound under the relative sampling error criteria (Definition 2.3). The result will also give a theoretical explanation to the observation made by Schaffer et al [21]. To begin with, let us first introduce some notation and preliminary results.

Definition 3.1 Let $f : S \to R^+$ be the fitness function. Given a population $\overrightarrow{X}= (X_1, \cdots, X_N)$, $X_i = (x_{i1}, \cdots, x_{il})^T$, $1 \leq i \leq N$, 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 $\overrightarrow{X}$ that have one or zero at the jth component (gene position) respectively, i.e.,

$$I_j^1 = \{1 \leq i \leq N; x_{ij} = 1\}, \quad I_j^0 = \{1 \leq i \leq N; x_{ij} = 0\}.$$

Write $F(\overrightarrow{X}) = \sum_{i=1}^N f(X_i)$ and $F_j^1(\overrightarrow{X}) = \sum\limits_{i \in I_j^1} f(X_i)$. We call

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

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

Lemma 3.1 Let $\mathcal{L}_j^1$ be the one order schema with the defining gene position j and let $p(\overrightarrow{X}, \cdot)$ be the sampling distribution given the current population $\overrightarrow{X}$. Then,

$$p(\overrightarrow{X}, \mathcal{L}_j^1) = a_j(\overrightarrow{X}) + (1 - 2a_j(\overrightarrow{X}))p_m,$$

where p_m is the mutation probability.

Proof: Similar to the proof of Theorem 3 in [8]. Q.E.D.

We now present a problem-dependent lower bound on the population size under the relative sample error criteria in Definition 2.3.

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 $f : S \to R^+$ be the fitness function. For a given population $\vec{X}$, let $\overline{p} = \min_{1 \le j \le l} \left(\frac{1}{2} - |\frac{1}{2} - p(\vec{X}, \mathcal{L}_j^1)| \right) > 0$ and assume that $\overline{p} > 0$. Then, any population size satisfying

$$N \geq N(\epsilon, \delta, l, \vec{X}, f) \triangleq \frac{1 - \overline{p} + \frac{1}{3}\epsilon}{2\overline{p}} \cdot \frac{1}{\epsilon^2}(\ln l + \ln \frac{4}{\delta}), \tag{19}$$

is strongly $(\epsilon, \delta, 1)$-effective, that is, for each $k \geq 1$ and $\vec{X}(k) = \vec{X}$, we have

$$P\left\{ \sup_{o(\mathcal{L})=1} \frac{1}{\sqrt{p(\vec{X}(k), \mathcal{L})}} \left| \frac{1}{N} N(\vec{X}(k+1), \mathcal{L}) - p(\vec{X}(k), \mathcal{L}) \right| \geq \epsilon \right\} < \delta. \tag{20}$$

Proof. To simplify the presentation, for a given population $\vec{X}$, we write

$$p_j^1 = p(\vec{X}, \mathcal{L}_j^1) \text{ and } p_j^0 = p(\vec{X}, \mathcal{L}_j^0).$$

For each $1 \leq i \leq N$ and $1 \leq j \leq l$, define

$$\xi_j^i = I_{\mathcal{L}_j^1}(X_i(k+1)).$$

Then, for each j, $(\xi_j^i, 1 \leq i \leq N)$ are conditionally independent given $\vec{X}(k) = \vec{X}$, and have the common Bernoulli distribution with the mean $E[\xi_j^i] = p_j^1$ and the variance

$$\sigma_j^2 = p_j^1 \cdot (1 - p_j^1).$$

For each j, it follows from Bernstein's inequality (See Lemma A.2 in Appendix A) that

$$P\left\{ \frac{1}{\sqrt{p_j^1}} \left| \frac{1}{N} N(\vec{X}(k+1), \mathcal{L}_j^1) - p_j^1 \right| > \epsilon \right\}$$

$$\leq P\left\{\left|\sum_{i=1}^{N}(\xi_j^i - p_j^1)\right| > N\epsilon p_j^1\right\}$$
$$\leq 2\exp\left[\frac{-\frac{1}{2}N\epsilon^2 p_j^1}{(1-p_j^1)+\frac{1}{3}\epsilon}\right]. \tag{21}$$

Similarly, we have

$$P\left\{\frac{1}{\sqrt{p_j^0}}\left|\frac{1}{N}N(\vec{X}\ (k+1), \mathcal{L}_j^0) - p_j^1\right| > \epsilon\right\}$$
$$\leq 2\exp\left[\frac{-\frac{1}{2}N\epsilon^2 p_j^0}{(1-p_j^0)+\frac{1}{3}\epsilon}\right]. \tag{22}$$

Now, suppose that the population size N satisfies (19), we get from (21) and (22) that,

$$P\left\{\sup_{o(\mathcal{L})=1}\frac{1}{\sqrt{p(\vec{X}\ (k), \mathcal{L})}}\left|\frac{1}{N}N(\vec{X}\ (k+1), \mathcal{L}) - p(\vec{X}\ (k), \mathcal{L})\right| \geq \epsilon\right\}$$
$$\leq \sum_{j=1}^{l} P\left\{\frac{1}{\sqrt{p_j^1}}\left|\frac{1}{N}N(\vec{X}\ (k+1), \mathcal{L}_j^1) - p_j^1\right| > \epsilon\right\}$$
$$+\sum_{j=1}^{l} P\left\{\frac{1}{\sqrt{p_j^0}}\left|\frac{1}{N}N(\vec{X}\ (k+1), \mathcal{L}_j^0) - p_j^0\right| > \epsilon\right\}$$
$$\leq 4l\exp\left[\frac{-\frac{1}{2}N\epsilon^2\overline{p}}{(1-\overline{p})+\frac{1}{3}\epsilon}\right]$$
$$\leq \delta. \tag{23}$$

So, N is a strongly $(\epsilon, \delta, 1)$-effective population size. The theorem is proved. Q.E.D.

Remark 3.1: In theorem 3.2, we have made an assumption that $\overline{p} > 0$ which is equivalent to $0 < p(\vec{X}, \mathcal{L}_j^1) < 1$ for all $1 \leq j \leq l$. If there is a schema (say $\mathcal{L}_j^1$) with $p(\vec{X}, \mathcal{L}_j^1) = 0$, we could simply exclude it from consideration when taking sup in (20). This is reasonable because if $p(\vec{X}, \mathcal{L}_j^1) = 0$, the sampling behavior in the schema $\mathcal{L}_j^1$ becomes deterministic and hence no sampling error exists. In practice, we could even only consider those schemata $\mathcal{L}_j$ such that $|p(\vec{X}, \mathcal{L}_j^1) - \frac{1}{2}|$ is smaller than a given threshold

α (say 0.1). In this case, the $\overline{p}$ in the lower bound (19) is greater than $\frac{1}{2} - \alpha$ so that the resulting lower bound is more practical.

We close this section by discussing the meanings of our lower bound in Theorem 3.2. Consider the term

$$\overline{p} = \min_{1 \leq j \leq l} (\frac{1}{2} - |p(\vec{X}, \mathcal{L}_j^1) - \frac{1}{2}|).$$

By Lemma 3.1, we have

$$\begin{aligned} |p(\vec{X}, \mathcal{L}_j^1) - \frac{1}{2}| &= |a_j^1(\vec{X}) - \frac{1}{2} - 2(a_j^1(\vec{X}) - \frac{1}{2})p_m| \\ &= |a_j^1(\vec{X}) - \frac{1}{2}| \cdot |1 - 2p_m| \\ &= 2|a_j^1(\vec{X}) - \frac{1}{2}| \cdot |p_m - \frac{1}{2}|. \end{aligned}$$

First, for a fixed mutation probability p_m, we see that the lower bound $N(\epsilon, \delta, l, \vec{X}, f)$ in Theorem 3.2 is an increasing function of $|a_j^1 - \frac{1}{2}|$. This implies that the larger the deviation of a_j^1 from $\frac{1}{2}$, the larger the necessary population size needed for GAs to have small sampling error. At this point, our lower bound is consistent with one of our previous results in [8], which states that the probability for gene loss to occur at the gene position j increases with $|a_j^1 - \frac{1}{2}|$.

Second, for fixed a_j^1's, since $\overline{p}$ is a decreasing function of $|p_m - \frac{1}{2}|$, we see that the necessary population size needed for GAs increases with $|p_m - \frac{1}{2}|$. This means that for $p_m \in (0, \frac{1}{2})$, the smaller the mutation probability, the larger the population size necessary for GAs to have good performance. Therefore, our result in Theorem 3.2 supports the experimental observation of Schaffer et al [21] that, to achieve good on-line performance, the population size is in an inverse relation to the mutation probability. Formally, we have the following Corollary.

Corollary 3.4 Assume that $0 < p_m < \frac{1}{2}$ and adopt the notations in Theorem 3.2. Then, any population size N with

$$N \geq N(\epsilon, \delta, p_m) = \frac{1 + \frac{1}{3}\epsilon - p_m}{2p_m} \cdot \frac{1}{\epsilon^2}\left(\ln l + \ln \frac{4}{\delta}\right), \tag{24}$$

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

Proof. The Corollary follows from Theorem 3.2 and the fact that for each j,

$$\frac{1}{2} - |p(\vec{X}, \mathcal{L}_j^1) - \frac{1}{2}| \geq \frac{1}{2} - (\frac{1}{2} - p_m) = p_m.$$

IV. Conclusion and Future Directions

In this paper, we have proposed two types of sampling error criteria for bounding the population size in genetic algorithms. Under the absolute sampling error criteria, a universal (problem-independent) lower bound on the population size was established. Based on the universal lower bound, the principle of implicit parallelism was interpreted from a new perspective. We have also established a problem-dependent lower bound on the population size which depicts how the necessary population size is related to the mutation probability and some population statistics. In particular, our result explains an experimental observation of Schaffer et al [21].

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 learning [25]. The work presented in this paper opens the door to using this powerful theory to analyze the behavior of genetic algorithms. This is perhaps not too surprising since the process of evolution is in essence a learning process.

In the future, we hope to sharpen the lower bounds on the population size so that they can be used in the practice of GAs. For example, using the tools developed by Haussler [25], it is possible to replace the term ϵ^2 in the lower bounds (10), (12), and (19) by ϵ. In this respect, any improvement of the Hoeffding's bounds such as those in [29] will be of help. Another promising direction for further research is to generalize the analyses in this paper to obtain lower bounds under criteria that minimize the sampling error uniformly over richer structures other than schemata, such as Radcliffe's formae [12].

It is also possible to use the ideas to bound the population sizes in real-coded genetic algorithms[27] and evolution strategies [28]. In such cases, the reasonable criteria should be those that minimize the sampling error uniformly over some kinds of structures such as the class of open sets, the class of rectangles, and the interval-schemata [27]. Since the number of these structures is infinite, more powerful tools such as Vapnik-Chervonenkis dimension(VC dimension) and the metric entropy [25], [26], may play an important role in the investigation.

Appendices

Appendix A. Two useful probabilistic inequalities

In the following, we present two probabilistic inequalities that are used in the proof of the main theorem. We refer the reader to Polland [24] for the detailed proof.

Lemma A.1(Hoeffding's Inequality) Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables with zero means and bounded ranges: $a_i \leq \xi_i \leq b_i$. Then, for each $\epsilon > 0$,

$$P\{|\xi_1 + \cdots + \xi_n| \geq \epsilon\} \leq 2 \exp \left[\frac{-2\epsilon^2}{\sum\limits_{i=1}^{n} (b_i - a_i)^2} \right].$$

Lemma A.2(Bernstein's Inequality) Let $\xi_1, \xi_2, \cdots, \xi_n$ be independent random variables with zero means and bounded ranges: $|\xi_i| \leq M$. Write σ_i^2 for the variance of ξ_i and suppose $V \geq \sigma_1^2 + \cdots + \sigma_n^2$. Then, for each $\epsilon > 0$,

$$P\{|\xi_1 + \cdots + \xi_n| \geq \epsilon\} \leq 2 \exp \left[\frac{-\frac{1}{2}\epsilon^2}{V + \frac{1}{3}M\epsilon} \right].$$

References

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

[2] D.E.Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning.* New York: Addison-Wesley,1989.

[3] X.Yao, " A review of evolutionary artificial neural networks," *International Journal of Intelligence Systems 8,* pp.539-567, 1993.

[4] D.B.Fogel, "An introduction to simulated evolutionary optimization," *IEEE Trans. on Neural Network*,vol.5,no.1,pp.3-14,1994.

[5] D.B.Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence,* New York: IEEE Press, 1995.

[6] G.Rudolph,"Convergence analysis of canonical genetic algorithms," *IEEE Trans. on Neural Networks*,vol.5,no.1, pp.96-101,1994.

[7] A.E.Eiben, E.H.L.Aarts, and K.M. van Hee, "Global convergence of genetic algorithms: A Markov chain analysis," in *Parallel Problem Solving from Nature*, H.P.Schwefel and R.Manner, Eds. Berlin: Springer-Verlag, pp.4-12, 1991.

[8] 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,*vol.8,no.5, pp.1165-1176,1997.

[9] C.C.Peck and A.P.Dhawan, " Genetic algorithms as global random search methods: an alternative perspective," *Evolutionary Computation,* vol.3,no.1,pp.39-80, 1995.

[10] M.D.Vose, “ Modelling simple genetic algorithms,” *Evolutionary Computation,* vol.3,no.4, pp.453-472, 1996.

[11] M.D.Vose and G.E.Liepins, “Schema disruption,” in *Proc. of the Fourth International Conference on Genetic Algorithms*, R.K.Belew and L.B.Booker,Eds. San Mateo, CA: Morgan Kaufmann, pp.237-242, 1991.

[12] N.J.Radcliffe, “The algebra of genetic algorithms,” *Annals of Mathematics and Artificial Intelligence 10,* pp.339-384, 1994.

[13] J.J.Grefenstette and J.E.Baker, “ How genetic algorithms work: a critical look at implicit parallelism,” in *Proc. of the Third International Conference on Genetic Algorithms*, J.D.Schaffer,Ed. San Mateo, CA: Morgan Kaufmann, pp.20-27, 1989.

[14] K.A.De Jong, *An Analysis of the Behavior of a Class of Genetic Adaptive Systems.* Doctoral Thesis, Department of Computer and Communication Sciences, University of Michigan, Ann Arbor,1975.

[15] D.Whitley, “ An executable model of a simple genetic algorithm,” in *Foundations of Genetic Algorithms.2*,L.D.Whitley, Ed. San Mateo,CA: Morgan Kaufmann,pp.45-62,1993.

[16] A.Bertoni and M.Dorigo, “ Implicit parallelism in genetic algorithms,” *Artificial Intelligence,* vol.61, pp.307-314, 1993.

[17] D.E.Goldberg, “ Sizing populations for serial and parallel genetic algorithms,” in *Proc. of the Third International Conference on Genetic Algorithms*, J.D.Schaffer,Ed. San Mateo, CA: Morgan Kaufmann, pp.70-79, 1989.

[18] D.E.Goldberg, K.Deb, and J.H.Clark,“ Accounting for noise in the sizing of populations,” in *Foundations of Genetic Algorithms-2,* D.Whitley, Ed. San Mateo, CA: Morgan Kaufmann Publishers, pp.127-140, 1993.

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

[20] 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 *Proceedings of the 1997 IEEE International Conference on Evolutionary Computation,* Piscataway: IEEE Press, pp.7-12, 1997.

[21] 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 *Proc. of the Third International Conference on Genetic Algorithms*, J.D.Schaffer,Ed. San Mateo, CA: Morgan Kaufmann, pp.51-60,1989.

[22] J.J.Grefenstette, “ Optimization of control parameters for genetic algorithms,” *IEEE Trans. on Systems, Man, and Cybernetics,* vol.16,no.1, pp.122-128, 1986.

[23] R.E.Smith and E.Smuda, “ Adaptively resizing populations: Algorithms, analysis, and first results,” *Complex Systems,* vol.9, pp.47-72, 1995.

[24] D.Pollard, *Convergence of Stochastic Processes,* New York: Springer-Verlag, 1984.

[25] D.Haussler, “Decision theoretic generalizations of the PAC model for neural net and other learning applications,” *Information and Computation,* vol.100, pp.78-150, 1992.

[26] A.Blumer, A.Ehrenfeucht, D.Haussler, and M.K.Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *Journal of the Association for Computing Machinery,*vol.36,no.4,pp.929-965,1989.

[27] L.J.Eshelman and J.D.Schaffer, " Real-coded genetic algorithms and interval-schemata," in *Foundations of Genetic Algorithms-2,* D.Whitley, Ed. San Mateo, CA: Morgan Kaufmann Publishers, pp.187-202, 1993.

[28] T.B*ä*ck, F.Hoffmeister, and H.-P.Schwefel, " A survey of evolution strategies," in *Proc. of the Fourth International Conference on Genetic Algorithms*, R.K.Belew and L.B.Booker, Ed.San Mateo, CA: Morgan Kaufmann, pp.2-9,1991.

[29] M.Talagrand, "The missing factor in Hoeffding's inequalities," *Ann.Inst.Henri Poincare, Probabilites et Statistiques,* vol. 31, no.4, pp.689-702, 1995.