
On The Supply Of Building Blocks

David E. Goldberg, Kumara Sastry, and Thomas Latoza

Illinois Genetic Algorithms Laboratory (IlliGAL)

Department of General Engineering

University of Illinois at Urbana-Champaign

104 S. Mathews Ave, Urbana, IL 61801

{deg,ksastry,latoza}@uiuc.edu

Abstract

This study addresses the issue of building-block supply in the initial population. Facetwise models for supply of a single building block as well as for supply of all schemata in a partition have been developed. An estimate for the population size required to ensure the presence of all raw building blocks has been derived using these facetwise models. The facetwise models and the population-sizing estimate are verified with computational results.

1 Introduction

The importance of building blocks (BBs) and their role in the working of GAs have long been recognized (Holland, 1975; Goldberg, 1989a). Elsewhere the problem of successful design of a selectorecombinative GA has been decomposed into six subproblems (Goldberg, Deb, & Clark, 1992): (1) Know what GAs are processing - building blocks (BBs), (2) ensure an adequate initial supply of raw BBs, (3) Ensure growth of superior BBs, (4) ensure the mixing of BBs, (5) ensure good decisions among competing BBs, and (6) solve problems with bounded BB difficulty. One of the essential steps towards successful design of a GA is making sure that the GA is well supplied with a sufficient number of the BBs required to solve a given problem. There are two approaches to address the BB supply question, a spatial and a temporal one. The spatial approach estimates the population size required to ensure diversity and the number of BBs present in the initial population. The temporal approach assumes the existence of a mutation or other diversity generator to return sufficient BB diversity on an appropriate time scale. In this study we restrict ourselves to BB supply in selec-

to-recombinative GAs, and hence address the spatial approach.

The objective of this study is to develop a facetwise model for supply of BBs and to estimate the population size required to guarantee the presence of all raw BBs in the initial population. Though, ensuring BB growth supersedes BB supply in the subsequent population, BB growth will be extremely difficult if BB supply is not ensured. While decision making governs population sizing usually, it is sometimes governed by BB supply. In such cases a facetwise model of BB supply is necessary for ensuring a successful GA design. It should be noted that there exist sophisticated models that combine the requirements of supply and decision making in a single model (Harik, Cantu-Paz, Goldberg, & Miller, 1997). These complex models are composed of simpler models, and blend the effects of those simple models. Therefore, developing simple facetwise models are useful for obtaining insight in more complex models. Here, we consider the BB supply question in isolation and derive a facetwise model using some straightforward probabilistic calculations. In this study we restrict ourselves to strings with fixed length, fixed alphabet cardinality, and fixed BB size. We start with a brief review of past work in this area. Then two facetwise models of the probability of BB supply are derived. Subsequently, an estimate of population size is presented by using the second facetwise model.

2 Brief Literature Review

A full review of past work on the supply of BBs is beyond the scope of this paper and hence a brief review is presented. Holland (1975) was the first to address the issue of BB supply. He estimated the number of BBs that receive at least a specified number of trials using Poisson distribution. A later study (Goldberg, 1989b) calculated the same quantity more exactly using bino-

mial distribution and used those calculations to study their effects on population sizing in serial and parallel computation. Recently Reeves (1993) proposed a population-sizing model for supply of BBs with fixed cardinality. However, he only considered BBs of unit size. Holland (1973, 1975) addressed the issue of decision making using the analogy of a two-armed bandit problem. De Jong (1975) incorporated the effects of noise in the decision process and proposed a population-sizing model based on the signal and noise characteristics of the problem. Goldberg and Rudnick (1991) developed a population-sizing model based on variance of fitness. A subsequent work (Goldberg, Deb, & Clark, 1992) developed a decision-based model to estimate population size to minimize decision errors. Their model is based on deciding correctly between the best and second best BBs in a partition in the presence of noise arising from other partitions. Harik, Cantu-Paz, Goldberg, and Miller (1997) refined this model by incorporating cumulative effects of decision making over a GA run. They also incorporated the effects of initial BB supply in their population-sizing model.

3 Facetwise Model for the Supply of a Single BB

Consider the probability of having a single k -position schema, p_k represented by one or more structures in a randomly generated population of size n . Let the population consist of strings of cardinality χ , k be the order of the schema. p_k is given by

$$p_k = 1 - \left[1 - \left(\frac{1}{\chi^k} \right) \right]^n. \quad (1)$$

Using the approximation, $(1 - r/n)^n \approx e^{-r}$, and recognizing that this approximation is sufficiently accurate even for modest values of n , we can write

$$p_k = 1 - \exp\left(-\frac{n}{\chi^k}\right). \quad (2)$$

The above equation is a simplified expression for the probability of having one or more successes at given schema. This model is compared to empirical results with alphabet cardinality of 2, 3, 4 and 5.

Figures 1(a), 1(b), 1(c) and 1(d) depict the proportion of runs out of 1000 trials for which at least one copy of a particular BB was present, for alphabet cardinalities of $\chi = 2$, $\chi = 3$, $\chi = 4$, and $\chi = 5$ respectively. The empirical results agree more accurately with the analytical results as the population size, n increases and as the term χ^k increases. An extension of this facetwise model to incorporate success at all schemas of a partition is presented in the next section.

4 Facetwise Model of Supply for Partition Success

When solving real-world problems, one does not have prior knowledge about a particular schema being superior to others in a partition. Hence it is necessary to ensure that all competing schemata in a partition are present. The decision process would then be able to consider all the relevant alternative schemata. Therefore in this section we extend the model developed in the previous section to ensure the presence of at least one copy of all the competing schemata in a partition. For example, if a particular problem requires the last two bits of a four-bit string to be evaluated jointly (**ff), we should ensure that all schemata in the partition (**00, **01, **10, **11) be present in the initial population.

We first derive an exact model to predict the probability of having at least one copy of all the competing schemata in a partition for the case $k = 1$ and alphabet cardinalities $\chi = 2, 3$, and 4. Then a generalized model of this probability for any schema-order and alphabet cardinality is presented. First considering the case $\chi = 2$, for a population of size n , the probability of having at least one copy of 1* and 0*, p_s can be written as: $p_s = 1 -$ probability of all the individuals being either 1* or 0*. That is,

$$\begin{aligned} p_s &= 1 - \left(\frac{1}{2^n} + \frac{1}{2^n} \right), \\ &= 1 - \frac{1}{2^{n-1}}. \end{aligned} \quad (3)$$

For $\chi = 3$, p_s is the complement of the probability that none of the individuals have at least one schema (0*, 1*, 2*). The number of possible ways of not having at least one of the schemas in a given partition, n_f , can be written as

$$n_f = 3 + 3 \sum_{i=1}^{n-1} \binom{n}{i}. \quad (4)$$

The first term represents the possibility of having identical schema in all individuals and the second term represents all possibilities of having two schemas. Using the binomial theorem, $\sum_{i=0}^n$

write $\sum_{i=0}^n \binom{n}{i} = 2^n$, we can

$$n_f = 3 + 3(2^n - 2) = 3(2^n - 1). \quad (5)$$

It can be easily seen that the total number of possible ways of schemas being present in the population is $\chi^n = 3^n$. The probability of success, p_s is given by

$$p_s = 1 - \frac{n_f}{3^n}$$

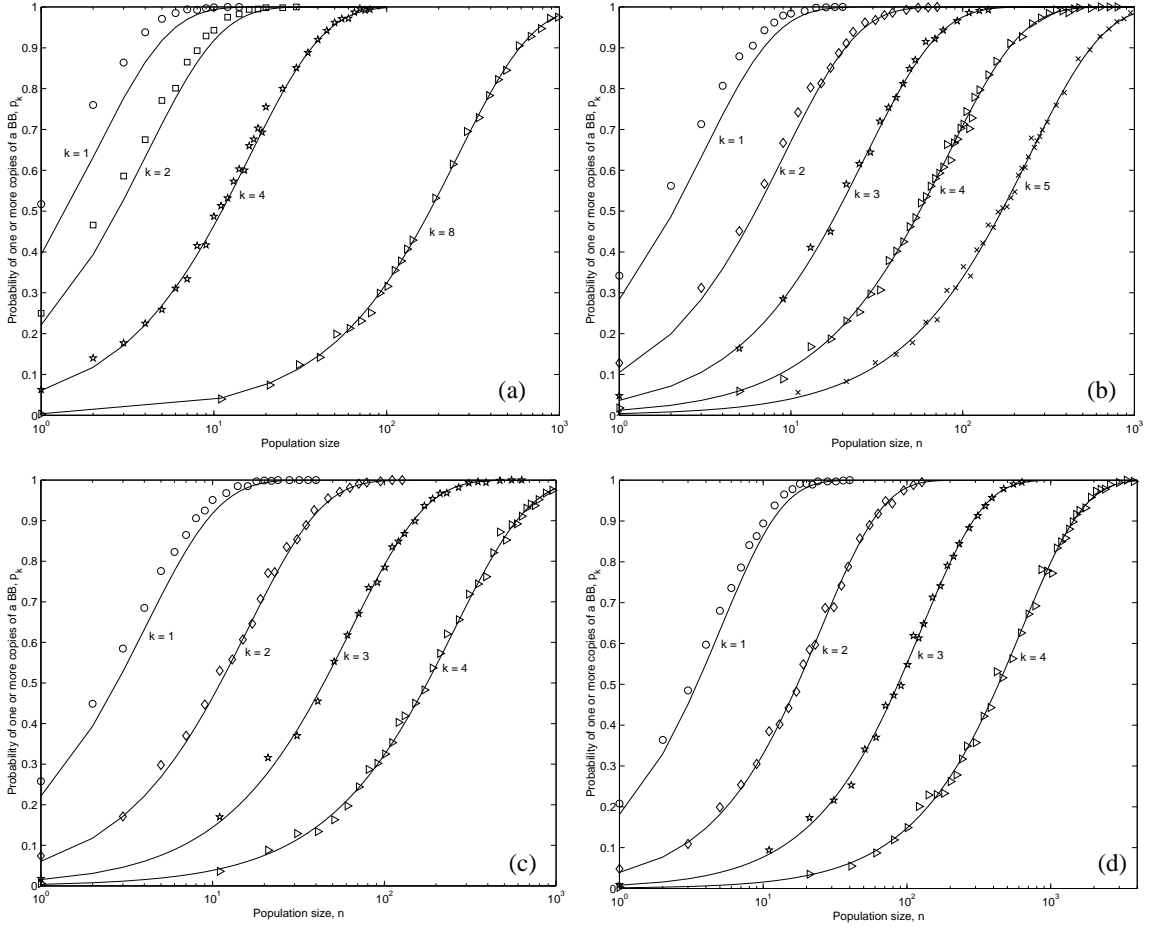


Figure 1: Verification of the facetwise model for a single BB supply (equation 2) with empirical results for different schema order values, k , and alphabet cardinalities, χ , as a function of population size, n . The empirical results depict the proportion of runs having at least one copy of a particular schemata out of 1000 trials. Cardinality of the string (a) $\chi = 2$, (b) $\chi = 3$, (c) $\chi = 4$, and (d) $\chi = 5$. Equation 1 matches the experimental results exactly and the agreement between equation 2 and experimental results increases with n and χ^k .

$$= 1 - \frac{2^n - 1}{3^{n-1}}. \quad (6)$$

Similarly, for $\chi = 4$, the number of possible ways of not having at least one of the schemata in a partition, n_f is given by

$$n_f = \binom{4}{1} + \binom{4}{2} \sum_{i=1}^{n-1} \binom{n}{i} + \binom{4}{3} \sum_{i=1}^{n-2} \left[\binom{n}{i} \sum_{k=1}^{n-i-1} \binom{n-i}{k} \right]. \quad (7)$$

Some straightforward simplification of the above equation yields,

$$n_f = 8 - 2^{n+1} + 8n + 4 \sum_{i=1}^{n-2} \left[\binom{n}{i} 2^{n-i} \right]. \quad (8)$$

Using the binomial theorem, we can equate

$$\sum_{i=1}^{n-2} \left[\binom{n}{i} 2^{n-i} \right] = 3^n - 2^n - 1 - 2n.$$

Substituting the above result in equation 8, we get

$$n_f = 4(3^n - 3 \cdot 2^{n-1} + 1). \quad (9)$$

The probability of success, p_s is given by

$$p_s = 1 - \frac{n_f}{4^n}, \\ = 1 - \frac{3^n - 3 \cdot 2^{n-1} + 1}{4^{n-1}}. \quad (10)$$

In general, for a schema-order value of k and alphabet cardinality of χ , the number of possible ways of not

having at least one of the schema in a given partition can be written as

$$n_f = \sum_{i=1}^{n_k-1} (-1)^{i-1} \binom{n_k}{i} (n_k - i)^n. \quad (11)$$

where $n_k = \chi^k$. Therefore the probability of success of having at least one copy of all schemata in a given partition is given by

$$p_s = 1 - \frac{1}{n_k^n} \left[\sum_{i=1}^{n_k-1} (-1)^{i-1} \binom{n_k}{i} (n_k - i)^n \right]. \quad (12)$$

We can easily see that when $n > n_k$, the above equation can be approximated as

$$p_s \approx 1 - n_k \exp\left(-\frac{n}{n_k}\right). \quad (13)$$

The exact model becomes complicated for higher values of schema order and does not give us useful insight. Hence we may derive a simpler form by assuming that the schema partition success values are independent. Then the probability of at least one success at each of the $n_k = \chi^k$ partitions, p_s is given by $p_s = p_k^{n_k}$. Using equation 2 we get

$$p_s = \left[1 - \exp\left(-\frac{n}{n_k}\right) \right]^{n_k}. \quad (14)$$

Again using the approximation, $(1 - r/n)^n \approx e^{-r}$, results in the following relation,

$$p_s = \exp\left[-n_k \exp\left(-\frac{n}{n_k}\right)\right]. \quad (15)$$

When $n > n_k$, using the approximation $(1 - x)^n \approx 1 - nx$ for small values of x , equation 14 can be written as

$$p_s \approx 1 - n_k \exp\left(-\frac{n}{n_k}\right). \quad (16)$$

Equations 13, and 16 show that the approximate model agrees with the exact model of the BB success probability for higher population sizes. The simplified model of equation 15 is compared to the exact result of equation 12 and empirical cases for $k = 1$ in figure 2. The empirical results match the exact model very accurately. The approximate model is a conservative estimate of the probability and it agrees well with empirical result at higher population sizes.

The above approximate model (equation 14) is verified with empirical results in figures 3(a)–(d). The plots compare the proportion of runs out of 1000 trials that have at least one copy of all members of a particular schema partition for alphabet cardinality of $\chi = 2, 3,$

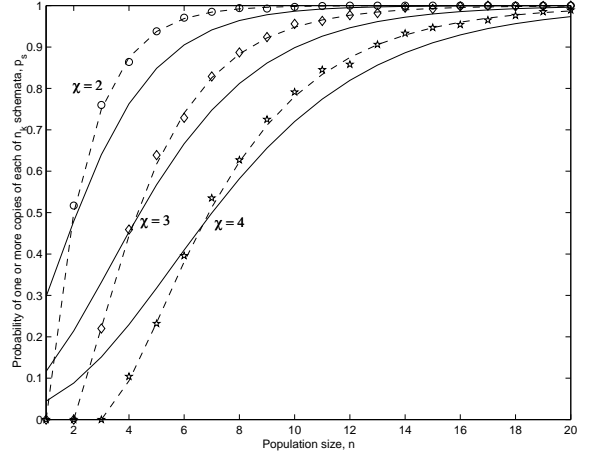


Figure 2: Comparison of the exact (equation 12) and approximate (equation 14) models for BB partition success with empirical results. The empirical results depict proportion of runs having at least one copy of all members of a particular schema partition out of 1000 trials for schema order value, $k = 1$ as a function of population size, n , for different alphabet cardinality, χ . The solid line represents equation 14, and the dashed line represents equation 12. The agreement between equations 12 and 14 increases as n increases.

4, and 5. As seen earlier we can see that the agreement between the empirical results and the analytical model increases with the increase in the schema order, k , and the population size, n .

5 Population Size for BB Supply

The facetwise model derived in the previous section will be rearranged in this section to estimate the population size required to ensure the presence of all BBs of a partition for problems of varying BB size, k , count, m , and alphabet cardinality, χ . Assuming that we can tolerate a probability α of not having all BBs in a given partition, and setting p_s to $1 - \alpha$, we can rewrite equation 15,

$$1 - \alpha = \exp\left[-n_k \exp\left(-\frac{n}{n_k}\right)\right]. \quad (17)$$

Taking logarithm on both sides and using the approximation $\log(1 - x) \approx -x$, for small values of x , gives

$$\alpha = n_k \exp\left(-\frac{n}{n_k}\right). \quad (18)$$

Solving the above equation for n yields

$$n = n_k (\log n_k - \log \alpha). \quad (19)$$

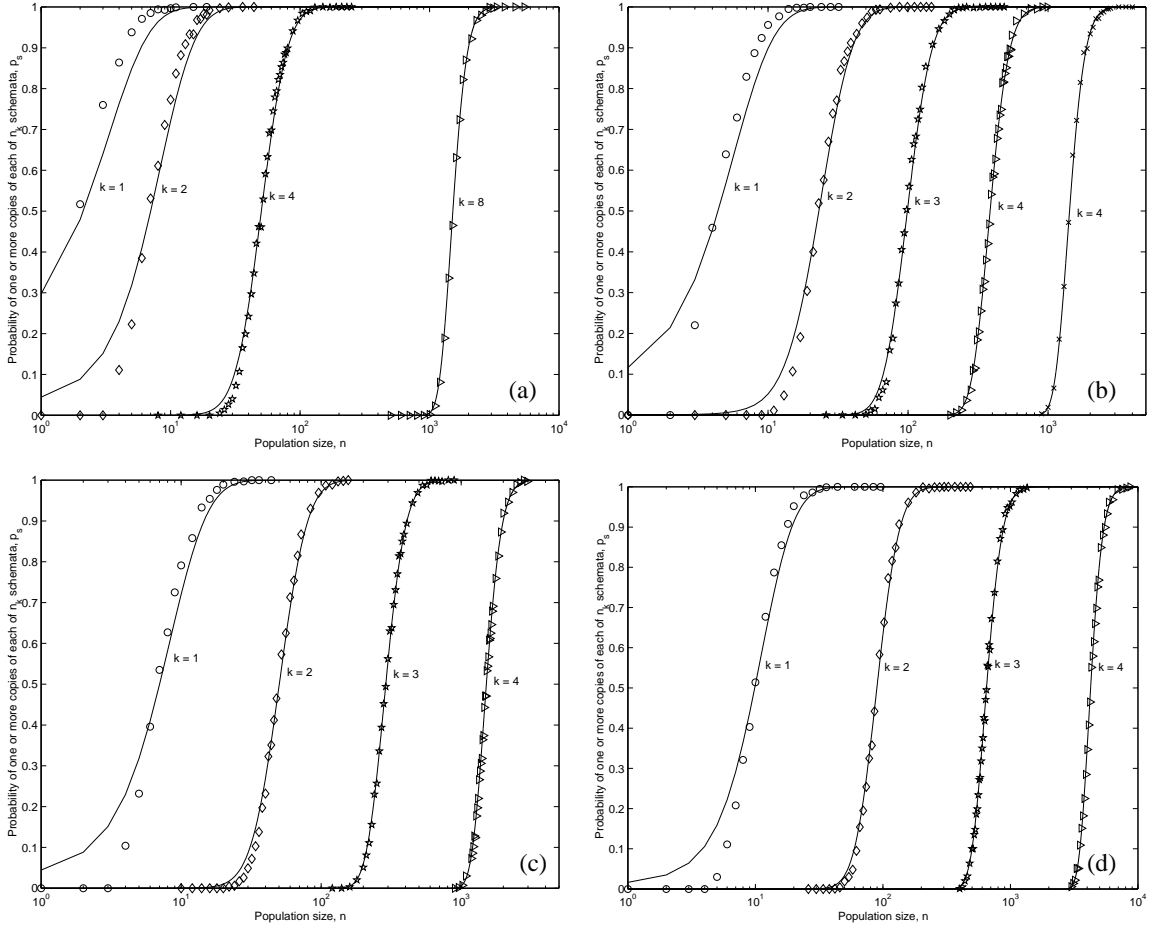


Figure 3: Verification of the model for BB partition success (equation 14) with empirical results for different schema order values, k , and alphabet cardinalities, χ , as a function of population size, n . The empirical results depict the proportion of runs having at least one copy of all members of a particular schema partition out of 1000 trials. Alphabet cardinality (a) $\chi = 2$, (b) $\chi = 3$, (c) $\chi = 4$, and (d) $\chi = 5$. Equation 12 matches the empirical results exactly and the agreement between equation 14 and experimental results increases with n and χ^k .

Using the definition of n_k we get,

$$n = \chi^k (k \log \chi - \log \alpha). \quad (20)$$

This relation yields the estimate of the population size required to ensure the presence of all BBs of a partition. We can simplify this relation further if we assume that the supply error is inversely proportional to the number of BBs, m , that is, $\alpha = 1/m$. Then the equation may be rewritten as

$$n = \chi^k (k \log \chi + \log m). \quad (21)$$

It can be easily seen that the above result can also be obtained from equations 13, and 16. The above population-sizing equation is an upper bound on the population size required to ensure the presence of all

the BBs in the initial population. This model is verified with empirical results in figures 4(a)–(d). The plots depict population size vs the number of building blocks, m , at different k values and different alphabet cardinality χ values. It can be easily seen from the plots that the model results agree with empirical results. The results shown are averaged over 100 independent runs, with all 100 runs yielding a supply error of less than or equal to $1/m$.

The population-sizing equation has two asymptotic cases. One when a problem is relatively large with respect to its complexity — when $m \gg \chi^k$ — and the other in which the problem is relatively complex with respect to its size — when $\chi^k \gg m$. In the first case, the population size required for ensuring the pres-

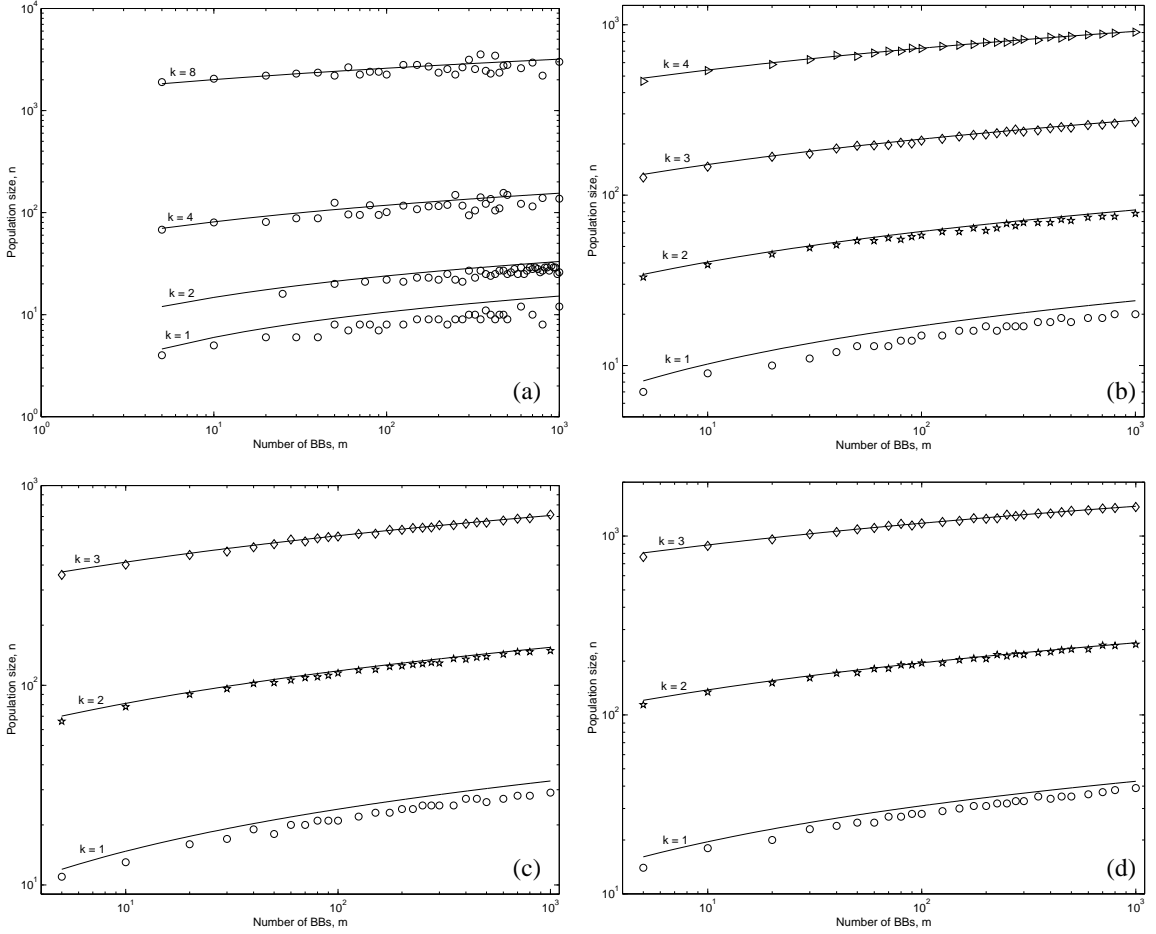


Figure 4: Verification of the population-sizing model for BB supply (equation 21) with empirical results for different partition sizes, k , and alphabet cardinalities, χ , as a function of number of BBs, m . The empirical results depict the population size required to represent all schemas in a partition of size k at a supply error of $1/m$. The experimental results are averaged over 100 runs. Alphabet cardinality (a) $\chi = 2$, (b) $\chi = 3$, (c) $\chi = 4$, and (d) $\chi = 5$. The agreement between equation 21 and experimental results increases with n and χ^k .

ence of all the schemas in a partition is $O(\chi^k \log m)$. In the later case, the population size required on BB supply grounds is $O(k\chi^k \log \chi)$. In both cases, the supply population-sizing is less than that required on decision-making grounds (Goldberg, Deb, & Clark, 1992; Harik, Cantu-Paz, Goldberg, & Miller, 1997).

6 Conclusions

In this paper, a detailed analysis of BB supply in the initial population, one of the six essential steps for a successful GA design, has been presented. Two facetwise models are derived, one for ensuring supply of a single schemata in a partition, and the other for ensuring supply of all competing schemata in a partition. The latter model has been employed to estimate

the population size required to ensure the presence of at least one copy of all raw BBs of a partition in the initial population. The population-sizing model indicates that for large easy problems the population size required on BB supply grounds is $O(\chi^k \log m)$, and for relatively complex problems the population size is $O(k\chi^k \log \chi)$.

Acknowledgments

This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0163. Research funding for this work was also provided by the National Science Foundation under grant DMI-9908252. Support was also provided by a grant from the U. S. Army Re-

search Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-2-0003. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, the U. S. Army, or the U. S. Government.

Thanks are also due to Ravi Srivastava and Abhishek Sinha for careful readings of this paper and for their comments.

References

- De Jong, K. (1975). *An analysis of the behavior of a class of genetic adaptive systems*. Doctoral dissertation, University of Michigan, Ann-Arbor, MI.
- Goldberg, D. (1989a). *Genetic algorithms in search optimization and machine learning*. Reading, MA: Addison-Wesley.
- Goldberg, D. (1989b). Sizing populations for serial and parallel genetic algorithms. In Schaffer, J. (Ed.), *Proceedings of the Third International Conference on Genetic Algorithms* (pp. 70–79). San Mateo, CA: Morgan Kaufmann.
- Goldberg, D., Deb, K., & Clark, J. (1992). Genetic algorithms, noise, and the sizing of populations. *Complex Systems*, 6, 333–362.
- Goldberg, D., & Rudnick, M. (1991). Genetic algorithms and the variance of fitness. *Complex Systems*, 5(3), 265–278.
- Harik, G., Cantu-Paz, E., Goldberg, D., & Miller, B. (1997). The gambler's ruin problem, genetic algorithms, and the sizing of populations. In Back, T., et al. (Eds.), *Proceedings of the IEEE International Conference on Evolutionary Computation* (pp. 7–12). Piscataway, NJ, USA: IEEE.
- Holland, J. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, 2(2), 88–105.
- Holland, J. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.
- Reeves, C. (1993). Using genetic algorithms with small populations. In *Proceedings of the Fifth International Conference on Genetic Algorithms* (pp. 92–99). San Mateo, CA: Morgan Kaufmann.