Population Sizing for Entropy-based Model Building in Discrete Estimation of Distribution Algorithms

Tian-Li Yu
Department of Electrical Engineering
National Taiwan University
No.1, Sec. 4, Roosevelt Rd., Taipei, Taiwan
tianliyu@cc.ee.ntu.edu.tw

Kumara Sastry
Illinois Genetic Algorithms Laboratory
University of Illinois at Urbana-Champaign
104 S. Mathews Ave., Urbana, IL 61801, USA
kumara@illigal.ge.uiuc.edu

David E. Goldberg
Illinois Genetic Algorithms Laboratory
University of Illinois at Urbana-Champaign
104 S. Mathews Ave., Urbana, IL 61801, USA
deg@uiuc.edu

Martin Pelikan
Missouri Estimation of Distribution Algorithms Laboratory
University of Missouri–St. Louis
One University Blvd., St. Louis, MO 63121, USA
pelikan@cs.umsl.edu

ABSTRACT

This paper proposes a population-sizing model for entropy-based model building in discrete estimation of distribution algorithms. Specifically, the population size required for building an accurate model is investigated. The effect of selection pressure on population sizing is also preliminarily incorporated. The proposed model indicates that the population size required for building an accurate model scales as $\Theta(m \log m)$, where m is the number of substructures of the given problem and is proportional to the problem size. Experiments are conducted to verify the derivations, and the results agree with the proposed model.

Categories & Subject Descriptors

G.1.6 [Mathematics of Computing]: Global Optimization–Analyze.

General Terms

Algorithms, Theory.

Keywords

Estimation of Distribution Algorithms, Genetic Algorithms, Population Sizing, Model Building, Entropy, Mutual Information.

1. INTRODUCTION

Genetic evolutionary computation (GEC) researchers have long realized the importance of population sizing on the success and efficiency of GEC. While using a smaller population usually yields low-quality solutions, using a population of size larger than required leads to wasting computational resources. Therefore, facetwise models, such as initial-supply [12] and decision-making models [11, 14], have been developed to model different bounds on population sizing for genetic algorithm (GA) success.

The issue of population sizing is equally critical, if not more, in estimation of distribution algorithms (EDAs) [17, 24], which build interaction models for the given problems and utilize the knowledge gained from the interaction models to efficiently recombine new solution candidates. For EDAs, the population should be sized properly not only to satisfy the initial supply and the need for making good decisions, but also to ensure the accuracy of the interaction model.

Pelikan, Sastry, and Goldberg [25] derived the population size required to build an accurate Bayesian model in the Bayesian optimization algorithm (BOA) to be

$$\Theta(m^{1.05}) \leq n \leq \Theta(m^{2.1}). \quad (1)$$

These bounds also apply to many other EDAs, and empirical findings show that n roughly scales as $\Theta(m^{1.4})$ [28]. However, a more refined model is required to explain the empirical findings and better understand population sizing. In addition, empirical findings also indicate that selection pressure affects population sizing and that an optimal selection pressure exists for model building.

Many different metrics have been used to detect interactions for model building. One of the most commonly used metrics is Shannon's entropy [29]. Typical examples for such entropy-based model building include ecGA [13], EBNA [6], BOA [22], the work of Wright [30], and DSMGA [32, 31].

The purpose of this paper is to develop a facetwise population-sizing model for entropy-based model building in EDAs. The model is anticipated to better explain the scalability of EDAs and capture the effect of selection pressure on population-sizing requirements.

This paper first derives the change of the entropy caused by selection, assuming an infinite population size. Then for a finite population size, the distributions of the sampled

GECCO'07, July 7–11, 2007, London, England, United Kingdom.

entropy are investigated. This paper then shows how selection pressure affects the distributions and subsequently population sizing. Finally, a population-sizing model based on decision making for entropy-based model building is derived, and experiments are conducted to verify the accuracy for the proposed model.

2. POPULATION SIZING

Facetwise and dimensional models have been very effective not only in the design of GAs, but also in understanding GA dynamics and mechanisms. Since our methodology depends on the facetwise models of population sizing, we briefly outline the models dictated by building-block (BB) supply, decision making, and accurate linkage learning in the remainder of this section. In addition, we also relate entropy to population sizing for the model building in EDAs.

2.1 BB Supply Model

The first step towards understanding population sizing is to tackle the issue of BB supply, where the minimum population size required to ensure the presence of at least one copy of all raw schemata is modeled. Holland [15] estimated the number of BBs that receive at least a specified number of trials using Poisson distribution. A later study [9] calculated the same quantity more exactly using binomial distribution and studied their effects on population sizing in serial and parallel computations. Reeves [26] proposed a population-sizing model for the supply of alphabets with a fixed cardinality. Recently, Goldberg *et al.* [12] developed facetwise models for ensuring BB supply in the initial population for GAs. They considered a population of fixed-length strings consisting of alphabets of cardinality χ and predicted that the population size required to ensure the presence of all competing BBs with a tolerance of $\epsilon = \frac{1}{m}$ is given by

$$n = \chi^k \left(k \log \chi + \log m\right), \tag{2}$$

where k is the order of a BB, and m is the total number of BBs.

2.2 Decision-Making Model

Goldberg *et al.* [11] proposed a population-sizing model based on decision making. The basic idea is that the population size should be large enough to alleviate sampling noises so a correct decision can be made between the correct BB and its most competing schema. If such correct decisions can be made for all BBs, then over generations, the global optima can be found by mixing all correct BBs. The model can be expressed as

$$n = c_1 2^k m \log m \frac{\sigma_{BB}^2}{d_{min}^2}, \tag{3}$$

where c_1 is a problem dependent constant, d_{min} is the minimal fitness difference between competing BBs, and σ_{BB}^2 is the fitness variance of a BB.

2.3 Gambler's Ruin Model

The decision-making model incorporates noises arising from other BBs. However, it assumes that if an incorrect decision is made in the first generation, GAs are unable to recover from the error. Harik *et al.* [14] refined the decision-making model by incorporating cumulative effects of decision making over time rather than in first generation only. They modeled the decision making between the correct BB and its most competing schema in a partition as a gambler's ruin problem. An approximated form of their population-sizing model is given by [14]:

$$n = \frac{\sqrt{\pi}}{2} \frac{\sigma_{BB}}{d_{min}} 2^k \sqrt{m} \log m. \tag{4}$$

The above equation assumes a failure probability $\alpha = \frac{1}{m}$.

2.4 Model-Building+Decision-Making Population Sizing

Facetwise modes for incorporating the effects of model building and BB-wise decision making on the population size have been analyzed for EDAs in general, and BOA and ecGA in particular [23, 25, 27, 28]. The population-sizing model that incorporates the effect of model building and its accuracy on population sizing of EDAs, and predicts the population size required to solve a problem with m BBs of order k with a failure rate of $\alpha = \frac{1}{m}$, is given by

$$n = c_2 \cdot 2^k \left(\frac{\sigma_{BB}}{d}\right)^2 m \log m, \tag{5}$$

where c_2 is another problem dependent constant.

2.5 Entropy and Population Sizing in EDAs

As mentioned in the introduction, one of the most commonly used metrics for model building in EDAs is entropy [19, 21]. Note that the derivations in this paper provide a *necessary* condition for accurate model building. For example, this paper focuses on a simple case with only two variables. Though BOA is able to handle dependencies between multiple variables, it starts to build the Bayesian network by considering only pair-wise dependencies at the very beginning. ecGA and DSMGA do not use only entropy; instead they adopt the minimum description length principle for model building. Nevertheless, the signals between dependent variables have to be significant enough for EDAs to detect regardless of the model complexity. In any case, this paper studies the minimal population size requirement that is large enough to alleviate sampling noises so that signals between two dependent variables are detectable for the EDA to build an accurate model.

3. THE ENTROPY CHANGE CAUSED BY SELECTION FOR INFINITE SAMPLING

We start to derive the population-sizing model by investigating the entropy metric, which is commonly used in many EDAs. Specifically, in this section, we investigate the difference of entropy of two genes before and after selection assuming an infinite population size. To make derivations feasible, we investigate only one generation. In other words, we investigate the population size needed for build an accurate model in the first generation. Although it is not necessary for a EDA to build such an accurate model in the first generation to solve the given problem, this assumption provides bounds for population sizing. In addition, later experiments show that the population size under this assumption is of the same order as the population size needed for solving the given problem for EDAs.

The loss in entropy by jointing two random variables together defines the mutual information: $\mathbb{I}(X;Y) = \mathbb{H}(X) +$

$\mathbb{H}(Y) - \mathbb{H}(X;Y)$ [5]. $\mathbb{H}$ is Shannon's entropy [29] and defined as $\mathbb{H}(X) = \mathbb{H}(\vec{p}) = -\Sigma_i p_i \log p_i$ for a discrete random variable X, where $\vec{p}$ is the occurrence probability vector of the events of X. For an entropy-based EDA to detect the existence of the interaction between genes X and Y, the sampled mutual information of X and Y needs to be significant enough.

The scenario in this paper is that after unbiased initialization of the population, the EDA performs binary tournament selection and then builds the interaction model. The population size required for building an accurate model is then investigated.

We believe that nearly decomposable problems represent a broad class of problems [10], and the success of EDAs in real-world application well supports the argument [17, 24]. The following derivation adopts a decomposable problem composed of the Royal road function [18]. The Royal road function serves as a worst case scenario for model building because given the minimal fitness difference d_{min}, the fitness differences between the best schema and all other $(2^k - 1)$ schemata are all d_{min}. In other words, the other $(2^k - 1)$ schemata in the Royal road function are equally competitive. Therefore, for a fixed d_{min}, the growth of the correct schema of the Royal road function is the slowest under a fixed selection pressure. Despite of the use of the Royal road function, our model is expected to be accurate for other nearly decomposable problems only with a different constant. For example, the experiments later in this paper show that the proposed model works on the trap function [8], which deceives hill-climbing algorithms.

To simplify the derivation, a bipolar Royal road function of order k is defined as follows to equalize the growth rates of 0 and 1 for every gene.

$$R_k(\vec{x}) = \begin{cases} 1 & \text{if } \vec{x} = \underbrace{111\cdots1}_{k} \\ 1 & \text{if } \vec{x} = \underbrace{000\cdots0}_{k} \\ 1-d & \text{otherwise.} \end{cases} \tag{6}$$

The derivation is based on decision making. Similar derivations can be found elsewhere [11, 25]. The fitness of an additively decomposable problem with m BBs is defined as

$$f(\vec{x}) = \Sigma_{i=0}^{m-1} R_k(x_{ik+1}x_{ik+2}\cdots x_{ik+k}). \tag{7}$$

Since the total number of BBs is proportional to the problem size, for simplicity, the terms "problem size" and "number of BBs" are interchangeable in the context of scalability for the rest of this paper.

Without loss of generality, the first two genes are chosen to calculate the mutual information between two dependent genes. Let X and Y be two random variables representing the first and second genes respectively. Also, for a quantity Q before selection, let Q' denote the same quantity after selection. Note that the bipolar Royal road function defined in Equation 6 is not biased to 0 or 1. For an infinite population size, at any given locus (the position of a gene), half of the population contains 0 while the other half contains 1 both before and after selection. Therefore, the entropy for an individual gene can be calculated as

$$\begin{aligned} &\mathbb{H}(X) = 1, \quad \mathbb{H}(X') = 1, \qquad \text{and} \\ &\mathbb{H}(Y) = 1, \quad \mathbb{H}(Y') = 1. \end{aligned} \tag{8}$$

The following derivation calculates the joint entropy of X and Y by investigating the competition between schemata concerning the first two genes. Define the following notation for schemata:

$$H_{xy} = xy\underbrace{**\cdots*}_{l-2}, \tag{9}$$

where x and y are 0 or 1, and l is the problem size. Define two sets $H_+ = \{H_{00}, H_{11}\}$ and $H_- = \{H_{01}, H_{10}\}$, and let F_+ and F_- be their corresponding fitness values:

$$F_+ = f(H_+) = f(H_{00}) + f(H_{11}). \tag{10}$$

$$F_- = f(H_-) = f(H_{01}) + f(H_{10}). \tag{11}$$

According to the central limit theorem [7], the distributions of F_+ and F_- can be approximated as Gaussian distributions when the population size is large. The variances of F_+ and F_-, defined as $\sigma^2_{F_+}$ and $\sigma^2_{F_-}$ respectively, are different but very close. By treating other $(m-1)$ BBs as external noises, these variances can be bounded and approximated as:

$$\begin{aligned} &(m-1)\sigma^2_{BB} \le \sigma^2_{F_+} \le m\sigma^2_{BB} \\ &(m-1)\sigma^2_{BB} \le \sigma^2_{F_-} \le m\sigma^2_{BB} \\ \Rightarrow \quad &\sigma^2_{F_+} \simeq \sigma^2_{F_-} = m\sigma^2_{BB} \cdot (1 - \mathcal{O}(\frac{1}{m})), \end{aligned} \tag{12}$$

where σ^2_{BB} is the fitness variance of a BB. The difference between those two variances is small and can be neglected when m is large.

Define $Z = F_+ - F_-$. Z is a normally distributed random variable with the following mean and variance.

$$E[Z] = \frac{d}{2^{k-2}}. \tag{13}$$

$$Var[Z] = \sigma^2_{F_+} + \sigma^2_{F_-} = 2m\sigma^2_{BB} \cdot (1 - \mathcal{O}(\frac{1}{m})). \tag{14}$$

The probability that H_+ wins over H_- in a binary tournament is given by $\Phi(\frac{E[Z]}{\sqrt{Var[Z]}})$, where Φ is the cumulative standard Gaussian probability density function. Define a decision variable z as:

$$z = \frac{E[Z]}{\sqrt{Var[Z]}} = \frac{d}{2^{k-2}\sqrt{2m} \cdot \sigma_{BB}} + \mathcal{O}(m^{-1.5}). \tag{15}$$

For a large m, z is small, and $\Phi(z)$ can be approximated by $\frac{1}{2} + \frac{z}{\sqrt{2\pi}} - \mathcal{O}(z^3)$ [1], which yields

$$\Phi(z) = \frac{1}{2} + \frac{d}{2^{k-1}\sqrt{\pi m} \cdot \sigma_{BB}} \pm \mathcal{O}(m^{-1.5}). \tag{16}$$

Define p_+ and p_- as the proportions of H_+ and H_- in the population, respectively. Before selection, $p_+ = p_- = \frac{1}{2}$. The proportions after selection can be calculated as:

$$p'_+ = p^2_+ + 2p_+ \cdot p_- \cdot \Phi(z) = \frac{1}{2} + \frac{\triangle_m}{4}, \tag{17}$$

$$p'_- = p^2_- + 2p_+ \cdot p_- \cdot \Phi(-z) = \frac{1}{2} - \frac{\triangle_m}{4}, \tag{18}$$

$$\text{where} \quad \triangle_m = \frac{d}{2^k\sqrt{\pi m} \cdot \sigma_{BB}}. \tag{19}$$

Equations 17 and 18 describe the changes of the proportions of H_+ and H_- caused by binary tournament selection.

Now calculate the joint entropy of the first two genes before and after selection:

$$\mathbb{H}(X;Y) = \mathbb{H}(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}) = 2. \quad (20)$$

$$\mathbb{H}(X';Y') = \mathbb{H}(\frac{p_+}{2},\frac{p_+}{2},\frac{p_-}{2},\frac{p_-}{2}) = 2 - \frac{\triangle_m^2}{8\ln 2} + \mathcal{O}(\triangle_m^4). \quad (21)$$

Combining Equations 8, 20, and 21, the change of the mutual information of the first two genes before and after selection is given by

$$\mathbb{I}(X';Y') - \mathbb{I}(X;Y) = \frac{\triangle_m^2}{8\ln 2} - \mathcal{O}(\triangle_m^4). \quad (22)$$

4. DISTRIBUTION OF ENTROPY FOR FINITE SAMPLING

This section investigates the effect of selection on the sampled entropy for a finite population. In the case of a finite population, the distribution of the sampled mutual information needs to be considered. Generally speaking, the selection operator increases the sampled mutual information between dependent genes and has no effect on independent genes. Specifically, this section derives the mean and variance of the sampled mutual information.

Define M_0 and M_1 as the mutual information after selection between pairs of independent genes and dependent genes, respectively. According to the unbiased initialization assumption, $M_0 = 0$, and M_1 is given by Equation 22.

Let $\hat{M}_{0,n}$ and $\hat{M}_{1,n}$ denote the sampled mutual information for M_0 and M_1 respectively, where n is the number of samples. For an infinite number of samples, $E[\hat{M}_{0,\infty}] = M_0 = 0$, and $E[\hat{M}_{1,\infty}] = M_1$. For a finite number of samples, the means and variances of the sampled mutual information can be derived as follows using Taylor expansion [16].

$$E[\hat{M}_{0,n}] = \frac{1}{2n\ln 2} + \mathcal{O}(\frac{1}{n^2}). \quad (23)$$

$$Var[\hat{M}_{0,n}] = \frac{1}{2n^2\ln 2} + \mathcal{O}(\frac{1}{n^3}). \quad (24)$$

$$E[\hat{M}_{1,n}] = \frac{\triangle_m^2}{8\ln 2} + \frac{1}{2n\ln 2} + \mathcal{O}(\frac{1}{n^2}) - \mathcal{O}(\triangle_m^4). \quad (25)$$

$$Var[\hat{M}_{1,n}] = \frac{1}{2n^2\ln 2} + \frac{\triangle_m^2}{4n\ln 2} - \mathcal{O}(\frac{\triangle_m^2}{n^2}) + \mathcal{O}(\frac{1}{n^3}). \quad (26)$$

In Equation 26, the first term dominates for a small n while the second term dominates for a large n.

To verify the above derivations, several experiments are conducted by first fixing the problem size and investigating the effect of different population sizes on the sampled mutual information. Equations 23 to 26 indicate that the sampled mutual information difference, $E[\hat{M}_{1,n} - \hat{M}_{0,n}]$, is virtually independent of n. $Var[\hat{M}_{0,n}]$ is inversely proportional to n^2. Roughly speaking, $Var[\hat{M}_{1,n}]$ is inversely proportional to n^2 for a small n and inversely proportional to n for a large n. The empirical results shown in Figure 1 support the derivation. The experiments are done by applying binary tournament selection to the (m,k)-trap [8] with m fixed at 10 and $k = 4$. The minimal fitness difference between competing BBs is 0.25. All results are averaged over 10000 independent runs.

Now the effect of different problem sizes on the sampled mutual information is investigated by fixing the population size and varying the problem size. Note that $\triangle_m$ is inversely proportional to $\sqrt{m}$ (Equation 19). Neglecting insignificant terms, Equations 23 to 26 suggest that the difference between the two means, $E[\hat{M}_{1,n} - \hat{M}_{0,n}]$, is inversely proportional to m. Likewise, the variance $Var[\hat{M}_{0,n}]$ is virtually independent of m. $Var[\hat{M}_{1,n}]$ is virtually independent of m when n is small while roughly inversely proportional to m when n is large. Again, the derivation agrees with the empirical results shown in Figure 2.

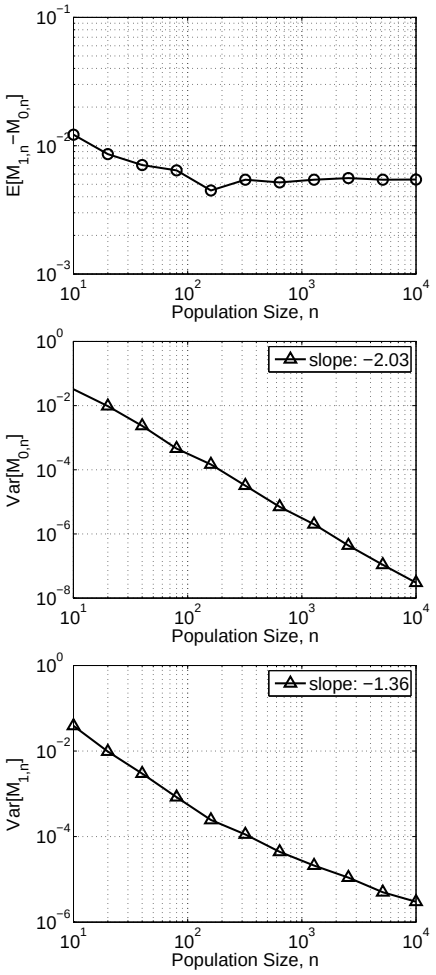


Figure 1: The effect of population size on the sampled mutual information. The problem size is fixed. The sampled mutual information difference is virtually independent of n. The sampled information variance for the pair of independent genes roughly scales $\Theta(n^{-2})$. That for the pair dependent genes scales closer to $\Theta(n^{-2})$ for small n and closer to $\Theta(n^{-1})$ for large n.

5. EFFECT OF SELECTION PRESSURE ON THE SAMPLED MUTUAL INFORMATION

This section extends the above analysis to tournament selection where the tournament size is s_{to}. While using results

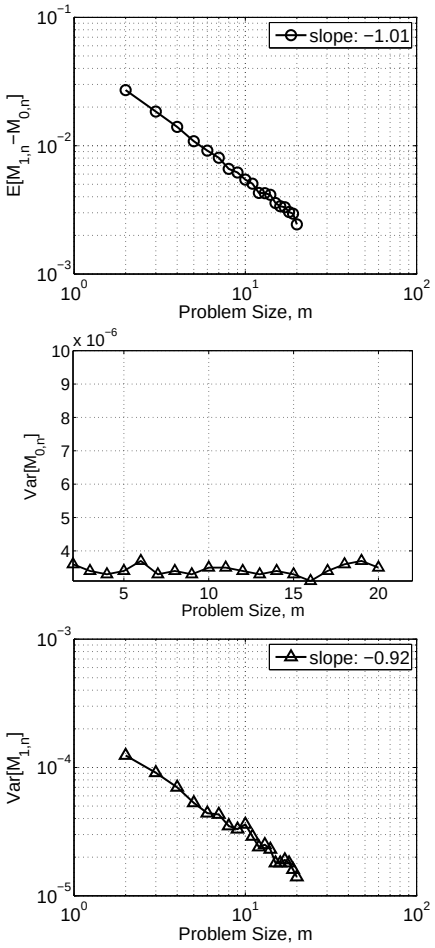


Figure 2: The effect of different problem sizes on the sampled mutual information. The population size is fixed. The sampled mutual information difference scales as $\Theta(m^{-1})$. The sampled information variance for the pair of independent genes is virtually independent of m. That for the pair of dependent genes scales as $\Theta(m^{-1})$.

from order statistics might accurately capture the effect of selection pressure on population sizing is acknowledged, this section approximates tournament selection by truncation selection to ease the analytical burden.

Consider the scenario where the selection operator is performed multiple times. It provides a similar effect of having an exponentially higher selection pressure. The statement is exactly true for truncation selection. In truncation selection, selecting the best half of the population twice results in exactly the same population as selecting the best quarter of the population.

Since all derivations in the previous section are based on tournament selection, a transformation from tournament selection to truncation selection is needed. Blickle and Thiele [4] gave the approximation of the selection intensity for tournament selection as $I = \sqrt{2\left(\ln s_{to} - \ln\sqrt{4.14\ln s_{to}}\right)}$. On the other hand, Bäck [3] approximated the selection intensity for truncation selection with a selection pressure s_{tr} as $I = s_{tr}\phi(\Phi^{-1}(1-\frac{1}{s_{tr}}))$, where ϕ is the probability density function of the standard Gaussian distribution and Φ is the cumulative density function.

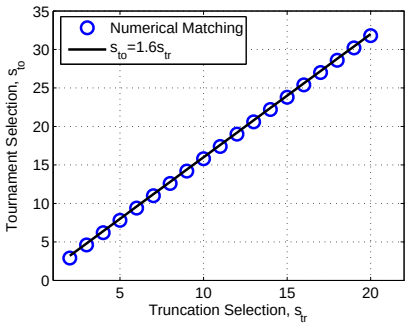


Figure 3: The relationship between the tournament size s_{to} and the selection pressure of truncation selection s_{tr} that yield the same selection intensity can be approximated as $s_{to} \simeq 1.6 s_{tr}$.

By setting the selection intensity to be same, s_{to} and s_{tr} can be solved numerically, and the following relation is obtained. (Figure 3).

$$s_{to} \simeq 1.6 s_{tr}. \tag{27}$$

Therefore, applying a binary tournament selection has a similar effect as applying truncation selection with a selection pressure $\frac{2}{1.6} \simeq 1.25$. Applying truncation selection t times results in a selection pressure $s_{tr} = 1.25^t$. When t is not too large, Equations 17 and 18 can be approximately modified as

$$p'_+ = \frac{1}{2} + \frac{t\triangle_m}{4} \qquad \text{and} \qquad p'_- = \frac{1}{2} - \frac{t\triangle_m}{4}. \tag{28}$$

As a result, the sampled information for a pair of dependent genes grows proportionally to t^2. On the other hand, the number of independent samples reduces from n to $\frac{n}{s_{tr}}$ after the selection procedure. Since the work in this paper focuses on the order of the relationship among population size, problem size, and selection pressure, all constants that are not related to any of these three factors are denoted as c_i for simplicity, where i distinguishes between different constants. Recall that $t = \frac{\ln s_{tr}}{1.25}$ and $s_{tr} = \frac{s_{to}}{1.6}$. The means and the variances of the sampled mutual information can be modeled as:

$$E[\hat{M}_{1,\frac{1.6n}{s_{to}}} - \hat{X}_{0,\frac{1.6n}{s_{to}}}] \simeq c_1(\ln\frac{s_{to}}{1.6})^2\triangle_m^2. \tag{29}$$

$$Var[\hat{M}_{0,\frac{1.6n}{s_{to}}}] \simeq c_2\frac{s_{to}^2}{n^2}. \tag{30}$$

$$Var[\hat{M}_{1,\frac{1.6n}{s_{to}}}] \simeq c_3\frac{(\ln\frac{s_{to}}{1.6})^2 s_{tr}\triangle_m^2}{n}. \tag{31}$$

Note that the approximation in Equation 31 neglects the first term in Equation 26, assuming n is large.

6. POPULATION SIZING FOR MODULARITY IDENTIFICATION

Previous sections model the means and variances of the sampled mutual information for a finite population. Uti-

lizing these models, this section derives a population-sizing model by the decision-making approach.

The distribution of the sampled mutual information can be approximated as a Gaussian distribution [16]. The decision-making error can be calculated as follows.

Define a variable τ as

$$\tau \triangleq \frac{E[Z]}{\sqrt{Var[Z]}} \simeq c_4 \frac{\ln \frac{s_{to}}{1.6}}{\sqrt{s_{to}}} \triangle_m \sqrt{n}, \tag{32}$$

where $Z = \hat{M}_{1,\frac{1.6n}{s_{to}}} - \hat{M}_{0,\frac{1.6n}{s_{to}}}$. The decision error ϵ is given by $1 - \Phi(\tau)$. For a large τ (small decision error), ϵ can be approximated as

$$\epsilon \simeq \frac{1}{\tau} e^{-\frac{\tau^2}{2}} \simeq \frac{c_5 \sqrt{s_{to}}}{\ln \frac{s_{to}}{1.6} \cdot \triangle_m \sqrt{n}} e^{-c_6 \frac{(\ln \frac{s_{to}}{1.6})^2 \triangle_m^2 n}{s_{to}}}. \tag{33}$$

For a problem with m BBs, there are mC_2^k pairs of dependent genes and $(C_2^{km} - mC_2^k)$ pairs of independent genes. Assuming that genes within a BB are maximally dependent, a BB can be treated as one decision variable, and only $C_2^m = \Theta(m^2)$ independent decisions need to be made correctly. Given the model accuracy to be $(1 - \frac{1}{m})$, the following relation holds.

$$(1 - \epsilon)^{\Theta(m^2)} \geq 1 - \frac{1}{m}. \tag{34}$$

For a small ϵ and a large m, Inequality 34 can be simplified as

$$\frac{1}{\epsilon} \geq \Theta(m^3). \tag{35}$$

With arithmetic manipulations, the following relation holds.

$$\ln(\frac{\ln \frac{s_{to}}{1.6} \cdot \triangle_m \sqrt{n}}{c_5 \sqrt{s_{to}}}) + c_6 \frac{(\ln \frac{s_{to}}{1.6})^2 \triangle_m^2 n}{s_{to}} \geq \Theta(\ln m). \tag{36}$$

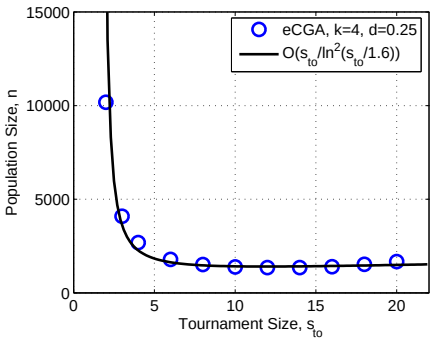
For a large n, the first term in Equation 36 can be neglected. By substituting $\triangle_m$ according to Equation 19, the following bound is obtained.

$$n \geq c_7 \frac{s_{to}}{(\ln \frac{s_{to}}{1.6})^2} 2^{2k} \frac{\sigma_{BB}^2}{d^2} m \ln m. \tag{37}$$

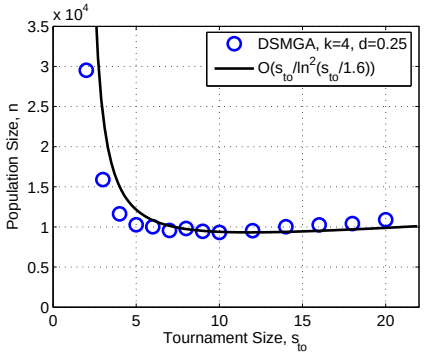
The population-sizing model given in Equation 37 differs from existing ones in three aspects. First of all, it incorporates the effect of selection pressure. Secondly, it scales as $\Theta(2^{2k})$ instead of $\Theta(2^k)$. Finally, it indicates the population size for model building should be $\Theta(m \ln m)$.

Figure 4 shows the relationship between the tournament size and the population size needed to build an accurate model with $(m - 1)$ BBs correctly identified. The results agree with the model qualitatively. Basically, for both smaller and larger s_{to}, a larger population size is needed to build an accurate model. Equation 37 also predicts a fixed optimal $s_{to}^* \simeq 11.8$. However, empirical results indicate that the optimal tournament size varies with the problem and the model-building procedure. This phenomenon is not yet captured in our model. The problem might lie in using truncation selection to approximate tournament selection. Even though the approximation ensures a similar selection intensity, in tournament selection, the number of copies of an individual is proportional to its rank, which is not the case for truncation selection.

The term 2^{2k} in Equation 37 is empirically verified. Figure 5 shows the relationship between the population size



(a) ecGA



(b) DSMGA

Figure 4: The relationship between the tournament size and the population size. Both the results and model indicates the existence of an optimal s_{to} around 10.

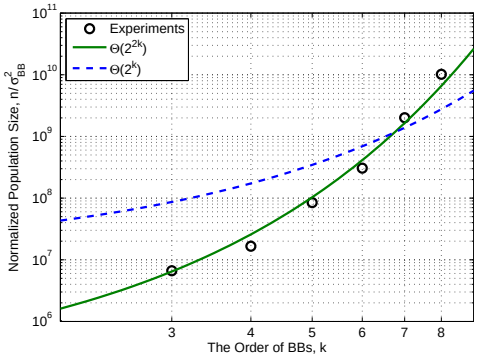
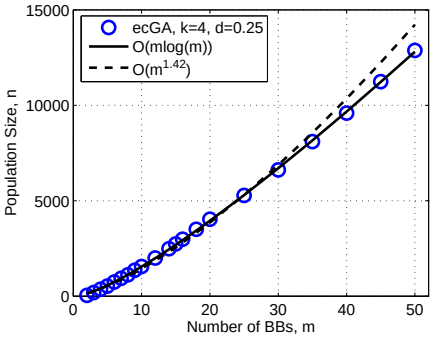


Figure 5: The relationship between the population size needed for DSMGA and the order of BBs, k. $\Theta(2^{2k})$ better describes the result than $\Theta(2^k)$ does.

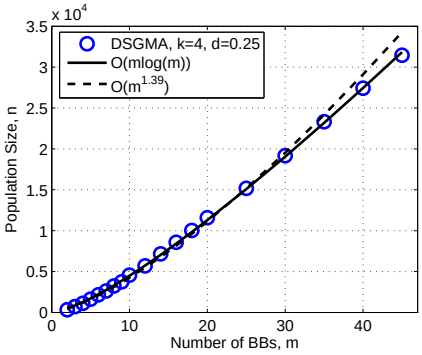
needed for the model building in DSMGA and the order of BBs, k, for an (m, k)-trap function, where $m = 10$. The minimal fitness difference between competing BBs is 0.1. As indicated in the figure, $\Theta(2^{2k})$ better describes the result than $\Theta(2^k)$.

Figure 6 shows the experimental results for ecGA and DSMGA on an (m, k)-trap function, where $k = 4$. The fitness difference between competing BBs is 0.25. The power-

law curve fitting is done by first-order polynomial fitting on the log-log scale. $\Theta(m \log m)$ provides a better description of the data than the power-law model.



(a) ecGA



(b) DSMGA

Figure 6: The scalability of the population size for different problem sizes. $n = \Theta(m \log m)$ is a better description of the results of both ecGA and DSMGA than the power-law model.

7. SUMMARY AND CONCLUSIONS

This paper presents a population-sizing model for entropy-based model building in EDAs. Specifically, the population size required for building an accurate model is investigated. The proposed model refines the required population size for model building from $\Theta(m^{1.05}) \leq n \leq \Theta(m^{2.1})$ to $n = \Theta(m \log m)$. It also corrects the term 2^k in existing population-sizing models to 2^{2k}. Those modifications are empirically verified. The proposed model also incorporates the effect of selection pressure on the population sizing requirements. Empirical results quantitatively agree with the proposed model for the scalability on the problem size. The modeling on selection pressure is qualitatively verified by experiments. To obtain a more accurate modeling on selection pressure, the derivation may need to utilize results from order statistics on the Gaussian distribution.

Compared with the existing population-sizing model for EDAs [25, 28], the proposed population-sizing model scales the same with the problem size as $\Theta(m \log m)$; they agree with each other. In other words, the population size required for building an accurate model is of the same order as that needed for EDAs to solve the problem. Recall that in EDAs, the population needs to be properly sized to satisfy the needs of BB supply, decision making, and model building. Among these three requirements, the population required for model building is usually large enough to satisfy the other two requirements for large-scale problems. Therefore, we can conclude that it is essential to build an accurate model for EDAs to solve large-scale problems.

Also, it is worth noting that the decision-making model [11] has a similar form. The difference is that in the decision-making model, decisions are made between competing BBs. Here, decisions are made between pairs of dependent and independent genes. The proposed model indicates that for a low selection pressure, the signal may not be strong enough to detect the existences of interactions; however, for a high selection pressure, sampling noises may cloud the signal. An optimal selection pressure exists somewhere in the middle for the model builder. Finally, although the proposed model is based on the entropy metric, a similar procedure should be applied to some other metrics such as nonlinearity [20] and simultaneity [2].

It has been shown that EDAs are capable of scalably solving many important real-world problems via problem decomposition [17, 24]. To enhance the optimization ability of EDAs or to even design a new EDA, it is important to study the behavior of EDAs on decomposable or nearly decomposable problems. The results of this paper are significant because this paper provides population-sizing bounds for model building that are applicable to a wide class of EDAs.

Finally, although this paper focuses on discrete EDAs, the methodology and derivations are not limited to EDAs only. More generally, this paper should also benefit the study of so-called *linkage-learning* problem [10] via entropy-based linkage model building.

ACKNOWLEDGMENTS

This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant FA9550-06-1-0096, the National Science Foundation under ITR grant DMR-03-25939 at Materials Computation Center, UIUC, and under CAREER grant ECS-0547013, and the University of Missouri in St. Louis through the High Performance Computing Collaboratory sponsored by Information Technology Services, and the Research Award and Research Board programs. 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, or the U.S. Government.

8. REFERENCES

[1] M. Abramowitz and L. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1970.

[2] C. Aporntewan and P. Chongstitvatana. Building-block identification by simultaneity matrix. *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1566–1567, 2003.

[3] T. Bäck. Generalized convergence models for tournament and (μ;λ) selection. *Proceedings of the Sixth International Conference on Genetic Algorithms (ICGA 1995)*, pages 2–8, 1995.

[4] T. Blickle and L. Thiele. A mathematical analysis of tournament selection. *Proceedings of the Sixth International Conference on Genetic Algorithms (ICGA'95)*, pages 9–16, 1995.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, pages 18–26. Wiley, New York, 1991.

[6] R. Etxeberria and P. Larrañaga. Global optimization using bayesian networks. *Proceedings of the Second Symposium on Artificial Intelligence Adaptive Systems*, pages 332–339, 1999.

[7] W. Feller. *An Introduction to Probability Theory*, volume 2. Wiley, New York, 1966.

[8] D. E. Goldberg. Simple genetic algorithms and the minimal, deceptive problem. In *Genetic Algorithms and Simulated Annealing*, chapter 6, pages 74–88. Pitman Publishing, London, 1987.

[9] D. E. Goldberg. Sizing populations for serial and parallel genetic algorithms. *Proceedings of the Third International Conference on Genetic Algorithms*, pages 70–79, 1989.

[10] D. E. Goldberg. *The design of innovation: Lessons from and for competent genetic algorithms*. Kluwer Academic Publishers, Boston, MA, 2002.

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

[12] D. E. Goldberg, K. Sastry, and T. Latoza. On the supply of building blocks. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pages 336–342, 2001.

[13] G. Harik. Linkage learning via probabilistic modeling in the ecga. IlliGAL Report No. 99010, University of Illinois at Urbana-Champaign, Urbana, IL, February 1999.

[14] G. Harik, E. Cantú-Paz, D. E. Goldberg, and B. L. Miller. The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Proceedings of the 1997 IEEE International Conference on Evolutionary Computation*, pages 7–12, 1997.

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

[16] M. Hutter and M. Zaffalon. Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis*, 48(3):633–657, 2005.

[17] P. Larrañaga and J. Lozano, editors. *Estimation of Distribution Algorithms*. Kluwer Academic Publishers, Boston, MA, 2002.

[18] M. Mitchell, S. Forrest, and J. H. Holland. The royal road for genetic algorithms: Fitness landscapes and GA performance. *Towards a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, pages 245–254, 1992.

[19] H. Mühlenbein and R. Höns. The estimation of distributions and the minimum relative entropy principle. *Evolutionary Computation*, 13(1):1–27, 2005.

[20] M. Munetomo and D. E. Goldberg. Identifying linkage groups by nonlinearity/non-monotonicity detection. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-1999)*, 1:433–440, 1999.

[21] J. Ocenasek. Entropy-based convergence measurement in discrete estimation of distribution algorithms. In L. Jose A., L. Pedro, and I. Inaki, editors, *Towards a New Evolutionary Computation: Advances in Estimation of Distribution Algorithms*, pages 39–49. Springer Verlag, New Yorks, 2006.

[22] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz. BOA: The Bayesian optimization algorithm. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-1999)*, I:525–532, 1999.

[23] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz. Bayesian optimization algorithm, population sizing, and time to convergence. *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 275–282, 2000.

[24] M. Pelikan, K. Sastry, and E. Cantú-Paz, editors. *Scalable Optimization via Probabilistic Modeling from Algorithms to Applications*. Springer, Berlin, 2006.

[25] M. Pelikan, K. Sastry, and D. E. Goldberg. Scalability of the Bayesian optimization algorithm. *International Journal of Approximate Reasoning*, 31(3):221–258, 2003.

[26] C. Reeves. Using genetic algorithms with small populations. *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 92–99, 1993.

[27] K. Sastry and D. E. Goldberg. On extended compact genetic algorithm. *Late-Breaking Paper at the Genetic and Evolutionary Computation Conference*, pages 352–359, 2000.

[28] K. Sastry and D. E. Goldberg. Designing competent mutation operators via probabilistic model building of neighborhoods. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2004)*, 2:114–125, 2004.

[29] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[30] A. Wright, R. Poli, C. Stephens, W. B. Landgon, and S. Pulavarty. An estimation of distribution algorithm based on maximum entropy. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2004)*, pages 343–354, 2004.

[31] T.-L. Yu and D. E. Goldberg. Conquering hierarchical difficulty by explicit chunking: Substructural chromosome compression. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2006)*, pages 1385–1392, 2006.

[32] T.-L. Yu, D. E. Goldberg, A. Yassine, and Y.-p. Chen. Genetic algorithm design inspired by organizational theory: Pilot study of a dependency structure matrix driven genetic algorithm. *Proceedings of Artificial Neural Networks in Engineering (ANNIE 2003)*, pages 327–332, 2003.