Stochastic Analysis of Schema Distribution in OneMax Problem

Hiroshi FURUTANI
Faculty of Engineering
University of Miyazaki
Miyazaki City, 889-2192

Susumu KATAYAMA
Faculty of Engineering
University of Miyazaki
Miyazaki City, 889-2192

Makoto SAKAMOTO
Faculty of Engineering
University of Miyazaki
Miyazaki City, 889-2192

Abstract

We have studied effects of stochastic fluctuation in the process of GA evolution of OneMax problem. We applied the Wright-Fisher model and the diffusion model for the analysis of GA with mutation and crossover. By using the diffusion model, we derived the stationary distribution of the first order schema frequency. In the comparison of numerical experiments and the theoretical calculations, we found that crossover is a very important factor determining the work of stochastic fluctuation.

1 Introduction

It is a difficult problem to determine the population size N in the application of GAs. Too small N may cause a poor performance in finding optimal solution(s), while too large N costs unnecessary computational power. The study of population sizing requires a stochastic treatment, which is in general a difficult task for the mathematical analysis.

In genetics, there are theories of Markov processes and diffusion equations. Population genetics uses the Wright-Fisher model of Markov processes [1], and diffusion models [2]. In the field of GAs, researchers apply Markov theory and its diffusion approximation. Nix and Vose proposed a stochastic model of GA evolution within the framework of Markov processes[3]. There are also analyses which use diffusion theories[4, 5].

Theoretical analyses in Genetics usually focus on one locus, and study the changes in frequencies of alleles within the locus. On the other hand, GA researchers treat multi-locus systems. This causes problems of dimensionality in Markov and diffusion approaches. To make the analysis tractable, Asoh and Mühlenbein restricted their study within the first order schemata[4]. In this study, we also used a mathematical framework of the first order schema theory, and investigated their evolution by using the Wright-Fisher model and diffusion equations. We analyzed the evolution of GA with mutation and crossover in OneMax problem.

2 Mathematical Model

A population is consisted of N individuals, and we fix N throughout the evolution process. We represent individuals by binary strings of length ℓ, and there are $n = 2^\ell$ genotypes. We identify integers and binary strings by

$$i \equiv < i(\ell), \cdots, i(1) >,$$

where $i(k) \in \{0, 1\}$ is the kth component of the binary string.

Denoting a genotype $< i(\ell), \cdots, i(1) >$ by i, we represent the number of individuals with genotype i at generation t by $N_i(t)$. We also use the relative frequency $x_i(t)$ for describing the evolution

$$x_i(t) \ = \ N_i(t)/N.$$

The process of fitness proportionate selection is given by

$$x_i(t+1) \ = \ f_i x_i(t)/\bar{f}(t), \tag{1}$$

where $\bar{f}(t)$ is the average fitness of the population

$$\bar{f}(t) \ = \ \sum_{i=0}^{n-1} f_i \, x_i(t). \tag{2}$$

3 Deterministic Model

We derive here the evolution equation for the first order schema within the framework of the deterministic theory.

3.1 Linkage Equilibrium

Linkage means the correlation between the different loci in a population, and if there is some correlation we call this linkage disequilibrium [6]. In GA applications, we usually generate an initial population by producing 0 and 1 randomly and independently at each bit position. The initial population of this type is obviously in linkage equilibrium. However, as the GA calculation proceeds, the population frequently goes into linkage disequilibrium state. The causes of this change are the functional form of the fitness and genetic drift in the selection process [6]. It should be noted that crossover and mutation have the effect of recovering linkage equilibrium, and if crossover works sufficiently, the population is in linkage equilibrium throughout the evolution process.

The distribution of a population in linkage equilibrium is given by using relative frequencies

$$x_i(t) \;=\; \prod_{k=1}^{\ell} h_{i(k)}(t), \tag{3}$$

where $h_{i(k)}(t)$ is a frequency of the first order schema at position k, and $i =< i(\ell), \ldots, i(1) >$. We also use the notation of $h_0^{(k)}$ and $h_1^{(k)}$ for the first order schema frequencies of bit 0 and bit 1, respectively.

3.2 OneMax Function

In this study, we use the OneMax fitness function f_i

$$f_i = \sum_{k=1}^{\ell} i(k). \tag{4}$$

Since the string of all ones $< 1, 1, \ldots, 1 >$ is the optimum solution for this landscape, bit 1 is the favorable allele at all positions.

In the deterministic schema theory, the evolution of the first order schema in linkage equilibrium is given by [8]. The relative frequency of the first order schema at position k is determined by

$$h_1^{(k)}(t+1) \;=\; a h_1^{(k)}(t) + b, \tag{5}$$

where constants a and b are

$$a \;=\; (1 - \frac{1}{\ell})\,(1 - 2\mu), \quad b \;=\; \frac{1}{\ell}(1 - 2\mu) + \mu.$$

The solution is given in terms of a,

$$b_0 \;=\; 1 - \frac{\mu}{2\mu + (1 - 2\mu)/\ell},$$

and the initial value $h_1(0)$

$$h_1(t) \;=\; a^t \left\{ h_1(0) - b_0 \right\} + b_0. \tag{6}$$

4 Markov Model

The stochastic model like Markov model explicitly treats the number of schemata. We consider the frequencies of two alleles at some locus. Two alleles are denoted by A and a, and the number of individuals having allele A and a are N_0 and N_1, respectively. Since $N = N_0 + N_1$ is constant, we consider N_1 in this analysis.

We first consider the case in which both alleles have the same fitness. The probability $Q(j|i)$ that $N_1 = i$ at generation t becomes $N_1 = j$ at generation $t+1$ is given by the binomial distribution

$$Q(j|i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}. \tag{7}$$

The probability $q_i(t)$ of $N_1 = i$ at generation t obeys the evolution equation

$$q_j(t+1) \;=\; \sum_{i=0}^{N} Q(j|i)\, q_i(t). \tag{8}$$

This equation is represented by using $(N+1)$-dimensional vector $\mathbf{q}(t)$

$$\mathbf{q}(t) \;=\; (q_0(t), \ldots, q_N(t))^T,$$

where T is transversion, and the matrix $Q = [Q(j|i)]$

$$\mathbf{q}(t+1) \;=\; Q\,\mathbf{q}(t),. \tag{9}$$

This model is called as the Wright-Fisher model[1].

We can express the evolution process analytically if we obtain the eigenvalues and eigenvectors of Q. We know the eigenvalues of the matrix [1]

$$1, \quad 1, \quad 1 - \frac{1}{N}, \quad \ldots, \prod_{i=1}^{N-1} \left(1 - \frac{i}{N}\right),$$

Unfortunately, we do not have the analytical expression of all eigenvectors. We only have two eigenvectors corresponding to the eigenvalue 1

$$\mathbf{q}_A \;=\; (1, 0, \ldots, 0)^T, \quad \mathbf{q}_a \;=\; (0, 0, \ldots, 1)^T,$$

These eigenvectors represent two absorbing states in Markov process. All processes converge to one of these absorbing states. The eigenvector $\mathbf{q}_A$ stands for the extinction state where there is no favorable allele in the population, and $\mathbf{q}_a$ for the fixation state of the favorable allele.

Next, we consider the case where two first order schemata have different fitness values. We use allele

A and a to denote binary values 0 and 1, respectively, and their fitness values are

$$f_A = 1, \quad f_a = 1 + s, \quad (s \geq 0).$$

The transition matrix $Q(j|i)$ is given by

$$\begin{aligned} Q(j|i) &= \binom{N}{j} u(s)^j \{1 - u(s)\}^{N-j}, \quad (10) \\ u(s) &= \frac{(1+s)i}{N+si}. \end{aligned}$$

This matrix also has a maximum eigenvalue 1 with multiplicity 2, and their corresponding eigenvectors are

$$\mathbf{q}_A = (1, 0, \ldots, 0)^T, \quad \mathbf{q}_a = (0, 0, \ldots, 1)^T.$$

It is also true that $\mathbf{q}_A$ and $\mathbf{q}_a$ are absorbing states.

Finally, we derive the transition matrix on the OneMax fitness. Replacing

$$h_1(t) \to i/N$$

in the evolution equation of the first order schema (5), we have the transition matrix

$$\begin{aligned} Q(j|i) &= \binom{N}{j} u(y)^j \{1 - u(y)\}^{N-j}, \quad (11) \\ u(y) &= ay + b = a\frac{i}{N} + b, \end{aligned}$$

where $y = i/N$. This transition matrix has eigenvalues $\nu_0, \ldots, \nu_N$

$$\nu_0 = 1, \ \nu_1 = a, \ldots, \nu_k = \prod_{i=0}^{k-1} a(1 - i/N), \ldots \quad (12)$$

These eigenvalues are independent of b, and the second largest eigenvalue $\nu_1 = a$ does not depend on N.

5 Diffusion Model

Though it is believed that Markov model can reproduce many evolution processes in biology, its mathematical analysis is very difficult. Therefore, the approximation of Markov model by the diffusion equation is used in genetics[1]. In the following, we derive the diffusion equation for OneMax problem.

We define

$$\Delta y(t) \ = \ y(t+1) - y(t),$$

Noting that i, therefore $y(t)$, is fixed, we calculate its expectation value

$$\mathrm{E}\{\Delta y(t)\} \ = \ \sum_{j=0}^{N} \frac{j}{N} \cdot q_j(t+1) - y(t),$$

and variance

$$\mathrm{V}\{\Delta y(t)\} \ = \ \frac{1}{N^2} \sum_{j=0}^{N} j^2 \cdot q_j(t+1) - \mathrm{E}\{\Delta y(t)\}^2.$$

Using

$$\sum_{j=0}^{N} \frac{j}{N} \cdot q_j(t+1) = \sum_{j=0}^{N} \frac{j}{N} \binom{N}{j} u(y)^j \{1 - u(y)\}^{N-j},$$

we have

$$\mathrm{E}\{\Delta y(t)\} \ = \ u(y) - y \ = \ (a-1)y + b. \quad (13)$$

In the same way, we have

$$\mathrm{V}\{\Delta y(t)\} \ = \ u(y)\{1 - u(y)\}/N. \quad (14)$$

From equations (13) and (14), we define two functions of y

$$M(y) \ = \ (a-1)y + b, \quad V(y) \ = \ u(y)\{1 - u(y)\}/N.$$

The diffusion approximation of Markov process in OneMax problem is given in terms of $M(y)$ and $V(y)$

$$\frac{\partial \phi(y,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial y^2}\{V(y)\phi(y,t)\} - \frac{\partial}{\partial y}\{M(y)\phi(y,t)\}. \quad (15)$$

This equation is called Kolmogorov forward equation, and $\phi(y,t)$ stands for the probability density function of the relative frequency y and time t.

We derive the solution of Kolmogorov forward equation (15). At $t \to \infty$, we try to obtain the stationary solution $\psi(y)$. Since $\partial\phi(y,t)/\partial t = 0$, $\psi(y)$ satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}y^2}\{V(y)\psi(y)\} - 2\frac{\mathrm{d}}{\mathrm{d}y}\{M(y)\psi(y)\} \ = \ 0.$$

By integrating the differential equation, we have

$$\psi(y) = C\,(ay+b)^{2c_1-1}\,(1-ay-b)^{2c_2-1} \quad (16)$$

where C is a normalization constant, and c_1 and c_2 are

$$c_1 = Nb/a^2, \quad c_2 = N(1-a-b)/a^2.$$

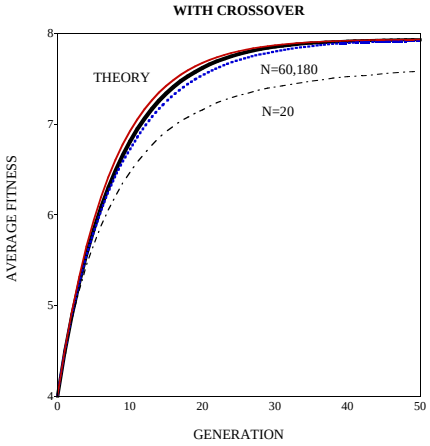

Figure 1: N dependence of $\bar{f}(t)$, with crossover.

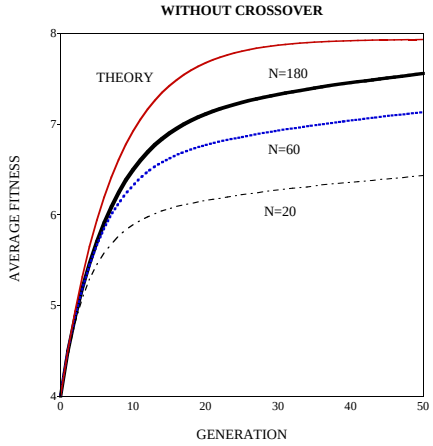

Figure 2: N dependence of $\bar{f}(t)$, without crossover.

6 Numerical Experiment

In this section, we compare results of the theoretical analysis with numerical experiments. We use the fitness proportionate selection and uniform crossover. The length of string is $\ell = 8$. We performed 1000 runs for each parameter set, and averaged over them.

Figure 1 shows the average fitness $\bar{f}(t)$ with the use of three population sizes, $N = 20, 60, 180$. We used the crossover rate $\chi = 1$ and mutation rate $\mu = 0.001$. The solid line with the label THEORY is obtained by the deterministic evolution equation (5) and $\bar{f}(t) = \ell h_1(t)$. Figure 2 shows the same results of GA calculations without crossover.

In the calculations with crossover, we observe small N dependence except for $N = 20$, and these results almost coincide with the deterministic theory. This fact can be explained by the fact that the population is in linkage equilibrium due to the action of crossover. In linkage equilibrium, Wright-Fisher model can describe the distribution of alleles in good approximation, in which the second largest eigenvalue does not have N dependence. Since the population size is small in the case of $N = 20$, there is large effect of genetic drift that makes the population in linkage disequilibrium.

On the contrary, we observe large differences in the results of different N in Fig.2. Especially, if N is small, the performance of the calculation is very poor. The strong N dependence suggests large effects of genetic drift in calculations without crossover.

Figure 3 explains the result of $N = 20$ with crossover. In 1000 repeated calculations, the numbers of fixations and extinctions of the favorable first order schema are shown as functions of generation. From this figure, we see that about 15% of favorable schema disappear from the population.

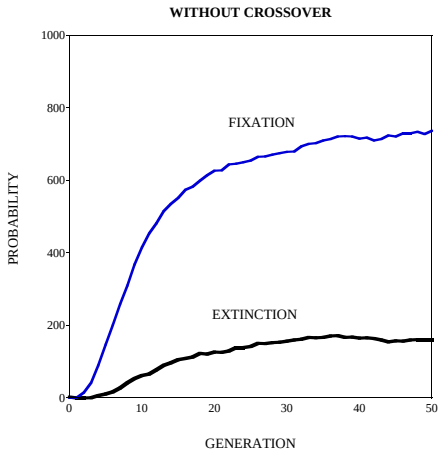

Figure 3: Fixation and extinction probabilities of the favorable first order schema. Without crossover. $N = 20$, $\ell = 8$, $\mu = 0.001$.

Linkage equilibrium due to crossover does not mean small N dependence. Figures 4 and 5 show distributions of the favorable first order schema at the stationary states ($t = 100$) with $\mu = 0.02$ and $\chi = 1$. The abscissa represents the relative frequency of bit one. The solid lines are the results of numerical calculations and dotted lines are those of diffusion approximation (16) . The large difference between two distributions comes from the difference in N.

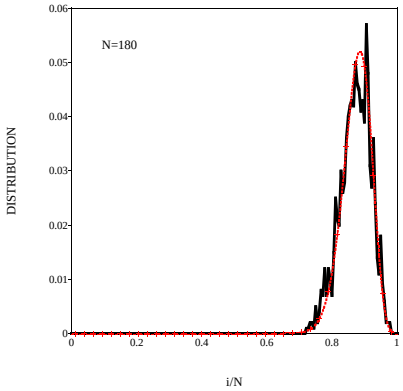

Figure 4: Distribution of the favorable first order schema, $\mu = 0.02$.

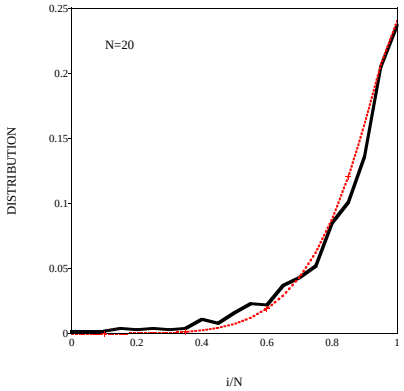

Figure 5: Distribution of the favorable first order schema, $\mu = 0.02$.

7 Summary

This study treats the schema evolution in OneMax by the use of the Wright-Fisher model and diffusion equations. If we can define favorable and unfavorable genes in one locus model, the fixation and extinction states are absorbing states in usual Markov processes, and this brings the N-dependence in GA calculations, for example, on the multiplicative landscape. While in the case of OneMax problem, the extinction state acts as a reflecting wall, and the probability of extinction state is zero. The reason is that the probability of the genotype $< 0, 0, \ldots, 0 >$ is zero. The assumption of linkage equilibrium means the probability of h_0 is also zero. Therefore the state $N_1 = 0$ cannot work as an absorbing wall. This partly explains the weak N-dependence in OneMax problem.

References

[1] Ewens, W.J.: Mathematical Population Genetics. I. Theoretical Introduction, Second Edition. Springer-Verlag, New York (2004)

[2] Crow, J. F., Kimura M.: An Introduction to Population Genetics Theory. Harper and Row, New York (1970)

[3] Nix, A. E., Vose, M. D.: Modelling Genetic Algorithm with Markov Chains. Annals of Mathematical and Artificial Intelligence. **5** (1992) 79–88

[4] Asoh, H., Mühlenbein, H.: On the Mean Convergence Time of Evolutionary Algorithms without Selection and Mutation. Parallel Problem Solving from Nature, Lecture Notes in Computer Science, **866**, Springer-Verlag, New York, (1994) 88-97

[5] Suzuki, H., Iwasa, Y.: Crossover Accelerates Evolution in GAs with a Babel-like Fitness Landscape: Mathematical Analyses. Evolutionary Computation, **7** (1999) 275-310; Errata: 8 (2000) 121-122

[6] Maynard Smith, J.: Evolutionary Genetics. 2nd edition, (Oxford University Press, Oxford, 1998).

[7] Furutani, H.: Schema Analysis of Genetic Algorithms on Multiplicative Landscape. in Recent Advances in Simulated Evolution and Learning, (World Scientific, Singapore, 2004) Chapter **6**

[8] Furutani, H.: Schema Analysis of OneMax Problem –Evolution Equation for First Order Schemata–. in Foundations of Genetic Algorithms 7, (Morgan Kaufmann, San Francisco, 2003) 9–26

[9] Whittaker, E.T. and Watson, G.N.: A Course of Modern Analysis, 4th Edition, (Cambridge University Press, Cambridge, 1927) 281–296