Estimating the Heritability by Decomposing the Genetic Drift*

Hideki Asoh
Electrotechnical Laboratory
Tsukuba, Ibaraki 305, Japan
asoh@etl.go.jp

Heinz Mühlenbein
GMD, Schloß Birlinghoven
D-53754 Sankt Augustin, Germany
muehlen@gmd.de

Abstract

Population genetics succeeded in recasting the macroscopic phenotypic Darwinian theory as formalized by Galton and Pearson into the microscopic genetic chance model initiated by Mendel. In order to do this the concept of genetic variance of a population was introduced. This concept and its application is one of the most difficult parts of population genetics. In this paper we define precisely how the genetic variance can be decomposed and how the method can be applied to haploid organisms. A fundamental theorem is proven which allows estimation of the heritability from microscopic genetic information of the population. It is indicated how the theorem can be used in the breeder genetic algorithm BGA.

1 Introduction

Genetics represents one of the most satisfying applications of statistical methods. Galton and Pearson found at the end of the last century a striking empirical regularity. On the average a son's height is halfway between that of his father and the overall average height for sons. They used data from about 1000 families. In order to see this regularity Galton and Pearson invented the scatter diagram, regression and correlation[4]. The regression coefficient between parent and offspring provides useful information for estimating the response of the population to selection[6].

Independently Mendel found some other striking empirical regularities like the reappearance of a recessive trait in one-fourth of the second generation hybrids. He conceived a chance model involving what are now called *genes* to explain his rules. He conjectured these genes by pure reasoning – he never saw any.

At first sight, the Galton–Pearson's results look very different from Mendel's, and it is hard to see how they can be explained by the same biological mechanism. Indeed Pearson wrote an article in 1904 claiming that his results cannot be derived from Mendel's laws. About 1920

*Technical report GMD-AS-TR-94-13

Fisher, Wright and Haldane more or less simultaneously recognized the need to recast the Darwinian theory as described by Galton and Pearson in Mendelian terms. They succeeded in this task, but unfortunately much of the original work is very abstruse and difficult to follow. The difficulty lies in the concept of *additive genetic variance* and its connection to *heritability*.

Several researchers in theoretical biology have tried to calculate the heritability and the correlation between relatives[2][3] using the Mendelian chance model. However, the cases analysed in the literature of quantitative genetics are closely related to diploid organisms. In the context of evolutionary algorithms, haploid virtual organisms are used most of the time. In this paper we will make the classical analysis more precise. From a mathematical point of view our analysis, while difficult, is better understandable. Furthermore we extend the analysis to haploid organisms using uniform crossover for recombination.

The key idea of the analysis is to decompose the genetic variance into an additive part and non-additive parts. We will derive a formula for calculating the covariance between midparent and offspring by the genetic variance. Some applications of the formula will be given. The outline of the paper is as follows. In section 2 some definitions and notations are given. In section 3 the fitness function is decomposed into an additive part and its interacting parts. For this decomposition several lemmas are proven, which will be used in section 4 for the main results. Because of the space limitations we only sketch some of the proofs. In section 5 some applications of estimating heritability by the fundamental theorem are given. Furthermore comparisons to experimental results are done. The importance of this theorem and its connection to Fisher's fundamental theorem of natural selection are discussed in section 6.

2 Regression and heritability

Consider a population of N haploid individuals. The generations are discrete and the size of the population is fixed at N. The mating scheme is random, that is, two parents are selected randomly from the population. The chromosomes of the parents are recombined giving an offspring. Self-fertilization is allowed. Neither selection nor mutation takes place. In the following discussion we assume that the population size N is large enough that the sample mean variance of interesting statistical values are near to theoretical mean and variance.

We assume that each individual has one chromosome of n loci. We denote the set of alleles for the i-th locus as Θ_i. Let the chromosomes of two parents be denoted by $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_n)$. Here $x_i, y_i \in \Theta_i$. Recombination is done by uniform crossover. The offspring $\mathbf{z} = (z_1, ..., z_n)$ is computed according to the following probability:

$$Prob[z_i = x_i] = \frac{1}{2}, \quad Prob[z_i = y_i] = \frac{1}{2}. \tag{1}$$

Recombination by uniform crossover is an adaptation of Mendel's chance model to haploid organisms.

Let the quantitative feature (fitness) of the individual with chromosome $\mathbf{x}$ be $f(\mathbf{x})$ and the relative frequency of the chromosome $\mathbf{x}$ in the population be $p(\mathbf{x})$. If we select two parents $\mathbf{x}$ and $\mathbf{y}$, then uniform crossover can generate 2^n offspring with equal probability of $1/2^n$. Note

that the possible chromosomes of the offspring depend on the chromosomes of the parents and uniform crossover. If for instance the alleles at some loci are equal, then this allele remains fixed in the chromosome of all possible offspring. We denote the chromosome of the i-th possible offspring by $\mathbf{z}_i$. Then the midparent–offspring covariance is given by

$$Cov_{\bar{p}o} = \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y}) \left(\frac{f(\mathbf{x}) + f(\mathbf{y})}{2} - \bar{f} \right) \frac{1}{2^n} \sum_{i=1}^{2^n} (f(\mathbf{z}_i) - \bar{f}). \tag{2}$$

Here $\bar{f}$ is the mean of the fitness value of the population,

$$\bar{f} = \sum_{\mathbf{x}} p(\mathbf{x}) f(\mathbf{x}). \tag{3}$$

We assume that mean and variance of the parent population are equal to the offspring population because there is no selection.

From the covariance the midparent–offspring correlation $Cor_{\bar{p}o}$ and the regression coefficient $b_{\bar{p}o}$ for midparent and offspring can be easily obtained. Using the well known formula $Var_{\bar{p}} = Var_p/2$ and the relation $Var_o = Var_p$ we get

$$Cor_{\bar{p}o} = \frac{\sqrt{2} Cov_{\bar{p}o}}{Var_p} \quad \text{and} \quad b_{\bar{p}o} = \frac{Cov_{\bar{p}o}}{Var_{\bar{p}}} = \sqrt{2} Cor_{\bar{p}o}. \tag{4}$$

Galton and Pearson called the regression coefficient the *heritability*. The heritability can be used to estimate the *response to selection* in breeding experiments.

In [8] the response to selection equation was introduced

$$R(t) = b_t S(t), \tag{5}$$

where $R(t) = M(t+1) - M(t)$ is called the response and $S(t) = M_s(t) - M(t)$ is called the *selection differential*. $M(t)$ and $M(t+1)$ are the mean fitness of generation t and $t+1$, and $M_s(t)$ is the mean fitness of selected parents.

The following theorem was proven in [9].

Theorem 1 *If the regression equation*

$$(f'_{ij} - \bar{f}) = b_{\bar{p}o}(\frac{f_i + f_j}{2} - \bar{f}) + \varepsilon_{ij}$$

with $E[\varepsilon_{ij}] = 0$ is valid between the fitness value of selected parents (f_i, f_j) and the offspring generated from them (f'_{ij}), then

$$R(t) = b_{\bar{p}o} S(t).$$

The above theorem connects the regression coefficient with the *realized heritability* $R(t)/S(t)$ in selection experiments. For applications of this theorem see [9]

In the following we will describe a method for calculating the regression coefficient from the microscopic information contained in the genetic chance model. The model will use the fitness function and the distribution of the genes in the population.

3 Decomposing the fitness function

In order to describe a formula for $Cov_{\bar{p}o}$, we decompose the fitness value f recursively. First we extract the constant term.

$$f(\mathbf{x}) = \alpha_0 + r_0(\mathbf{x}). \tag{6}$$

If $\alpha_0 = \bar{f}$ then $R_0 = \sum_{\mathbf{x}} p(\mathbf{x}) r_0(\mathbf{x})^2$ is minimized.

Next we extract the first order (additive) terms $\alpha_{(i)}(x_i)$ from the residual $r_0(\mathbf{x})$. Each $\alpha_{(i)}(x_i)$ depends only on the value of the i-th locus.

$$r_0(\mathbf{x}) = \sum_{i=1}^{n} \alpha_{(i)}(x_i) + r_1(\mathbf{x}). \tag{7}$$

If $\alpha_{(i)}(x_i)$ is equal to the conditional mean of $r_0(\mathbf{x})$, that is,

$$\alpha_{(i)}(x_i) = \sum_{\mathbf{x}|x_i} p(\mathbf{x}|x_i) r_0(\mathbf{x}) = \sum_{\mathbf{x}|x_i} p(\mathbf{x}|x_i) f(\mathbf{x}) - \bar{f},$$

then $R_1 = \sum_{\mathbf{x}} p(\mathbf{x}) r_1(\mathbf{x})^2$ is minimized. We denote this optimal value for $\alpha_{(i)}(x_i)$ as $f_{(i)}(x_i)$. Here $\sum_{\mathbf{x}|x_i}$ means that the allele of the i-th locus is fixed and the summation is taken over all variations of alleles at other loci. By $p(\mathbf{x}|x_i)$ we denotes the conditional probability induced from $p(\mathbf{x})$ under the condition of x_i being fixed. Exactly speaking this conditional probability should be noted as $p(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n|x_i)$. However in order to save the space we use the above notation. Note that $\sum_{\mathbf{x}|x_i} p(\mathbf{x}|x_i) = 1$ holds. The above decomposition is similar to first order schemata [5]. $f_{(i)}(x_i)$ is the average fitness of genotypes which belong to schemata $(*, ..., *, x_i, *, ..., *)$.

If $r_1(\mathbf{x}) \not\equiv 0$, we can proceed further to extract second order terms from $r_1(\mathbf{x})$:

$$r_1(\mathbf{x}) = \sum_{\substack{(i,j) \\ i<j}} f_{(i,j)}(x_i, x_j) + r_2(\mathbf{x}). \tag{8}$$

Here,

$$\begin{aligned} f_{(i,j)}(x_i, x_j) &= \sum_{\mathbf{x}|x_i,x_j} p(\mathbf{x}|x_i, x_j)\ r_1(\mathbf{x}) \\ &= \sum_{\mathbf{x}|x_i,x_j} p(\mathbf{x}|x_i, x_j)\ f(\mathbf{x}) - f_{(i)}(x_i) - f_{(j)}(x_j) - \bar{f}. \end{aligned}$$

As before, $f_{(i,j)}(x_i, x_j)$ minimizes $R_2 = \sum_{\mathbf{x}} p(\mathbf{x}) r_2(\mathbf{x})^2$ When we have n loci, we can iterate this procedure $n-1$ times and finally we get the decomposition of f

$$\begin{aligned} f(\mathbf{x}) &= \bar{f} + \sum_i f_{(i)}(x_i) + \sum_{(i,j)} f_{(i,j)}(x_i, x_j) + \cdots \\ &+ \sum_{\substack{(i_1,...,i_{n-1}) \\ i_1<...<i_{n-1}}} f_{(i_1,...,i_{n-1})}(x_{i_1}, ..., x_{i_{n-1}}) + r_{n-1}(\mathbf{x}). \end{aligned}$$

We will prove the following lemma by induction.

Lemma 1 *For all $k = 0$ to $n-1$*

$$\sum_{\mathbf{x}} p(\mathbf{x})\, r_k(\mathbf{x}) = 0.$$

Furthermore for all $k = 1$ to $n-1$ and for all combinations of $(i_1, ..., i_k)$, $i_1 < ... < i_k$

$$\sum_{x_{i_1},...,x_{i_k}} p_{(i_1,...,i_k)}(x_{i_1}, ..., x_{i_k})\, f_{(i_1,...,i_k)}(x_{i_1}, ..., x_{i_k}) = 0,$$

Here $p_{(i_1,...,i_k)}$ is a marginal probability distribution of $p(\mathbf{x})$ on $(x_{i_1}, ..., x_{i_k})$, that is,

$$p_{(i_1,...,i_k)}(x_{i_1}, ..., x_{i_k}) = \sum_{\mathbf{x}|x_{i_1},...,x_{i_k}} p(\mathbf{x}).$$

Proof Let us assume that the lemma is correct for $k = l$.

$$\sum_{\mathbf{x}} p(\mathbf{x})\, r_l(\mathbf{x}) = 0.$$

Using the definition of the conditional probability

$$p_{(i_1,...,i_l)}(x_{i_1}, ..., x_{i_l})\, p(\mathbf{x}|x_{i_1}, ..., x_{i_l}) = p(\mathbf{x}),$$

we obtain for $k = l+1$

$$\begin{aligned}
&\sum_{x_{i_1},...,x_{i_{l+1}}} p_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}})\, f_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}}) \\
&= \sum_{x_{i_1},...,x_{i_{l+1}}} p_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}}) \sum_{\mathbf{x}|x_{i_1},...,x_{i_{l+1}}} p(\mathbf{x}|x_{i_1}, ..., x_{i_{l+1}})\, r_l(\mathbf{x}) \\
&= \sum_{\mathbf{x}} p(\mathbf{x})\, r_l(\mathbf{x}) = 0.
\end{aligned}$$

Using this result we obtain for $k = l+1$

$$\begin{aligned}
&\sum_{\mathbf{x}} p(\mathbf{x})\, r_{l+1}(\mathbf{x}) = \sum_{\mathbf{x}} p(\mathbf{x})\, (r_l(\mathbf{x}) - \sum_{\substack{(i_1,...,i_{l+1}) \\ i_1<...<i_{l+1}}} f_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}})) \\
&= - \sum_{\substack{(i_1,...,i_{l+1}) \\ i_1<...<i_{l+1}}} \sum_{\mathbf{x}} p(\mathbf{x})\, f_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}}) \\
&= - \sum_{\substack{(i_1,...,i_{l+1}) \\ i_1<...<i_{l+1}}} \sum_{x_{i_1},...,x_{i_{l+1}}} \sum_{\mathbf{x}|x_{i_1},...,x_{i_{l+1}}} p(\mathbf{x})\, f_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}}) \\
&= - \sum_{\substack{(i_1,...,i_{l+1}) \\ i_1<...<i_{l+1}}} \sum_{x_{i_1},...,x_{i_{l+1}}} p_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}})\, f_{(i_1,...,i_{l+1})}(x_{i_1}, ..., x_{i_{l+1}}) = 0.
\end{aligned}$$

For $k = 0$

$$\sum_{\mathbf{x}} p(\mathbf{x})\, r_0(\mathbf{x}) = 0$$

obviously holds. ∎

Now we formulate several lemmas which are needed for the final theorem. For all of them the assumption of independence of each locus is needed. This condition is called *linkage equilibrium* in the literature of genetics. That is, it is assumed that $p(\mathbf{x})$ is given by

$$p(\mathbf{x}) = \prod_{i=1}^{n} p_i(x_i). \tag{9}$$

Here $p_i(x_i)$ denotes the relative frequency of allele x_i at the i th locus.

The next lemma can be proven by induction similar to lemma 1.

Lemma 2 *For all $k = 1$ to $n-1$, for all $l = 0$ to k, for all combination of $(j_1, ..., j_l)$, $j_1 < ... < j_l$ and for all fixed values of x_{j_s},*

$$\sum_{\mathbf{x}|x_{j_1},...,x_{j_l}} p(\mathbf{x})\, r_k(\mathbf{x}) = 0.$$

For all $k = 2$ to $n-1$, for all $l = 0$ to $k-1$, for all combination of $(j_1, ..., j_l) \subset (i_1, ..., i_k)$, $j_1 < ... < j_l$ and for all fixed values of x_{j_s},

$$\sum_{\mathbf{x}|x_{j_1},...,x_{j_l}} p(\mathbf{x})\, f_{(i_1,...,i_k)}(x_{i_1}, ..., x_{i_k}) = 0.$$

Lemma 3 (Orthogonality)

All terms in the decomposition of $f(\mathbf{x})$ are orthogonal to each other, that is,

$$\sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})\, f_{(i_1,...,i_k)}(z_{i_1}, ..., z_{i_k})\, f_{(j_1,...,j_l)}(z_{j_1}, ..., z_{j_l}) = 0 \tag{10}$$

hold unless $k = l$ and $(i_1, ..., i_k) = (j_1, ..., j_l)$ and $z_{i_t} = z_{j_t}$ for all t. For any $k = 1$ to $n-1$

$$\sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})\, f_{(i_1,...,i_k)}(z_{i_1}, ..., z_{i_k})\, r_{n-1}(z_1, ..., z_n) = 0 \tag{11}$$

hold. Here z_{i_t} is x_{i_t} or y_{i_t}, z_{j_t} is x_{j_t} or y_{j_t} and z_t is x_t or y_t.

Proof Without loss of generality we can assume that $k \geq l$, $z_{i_1} \notin \{z_{j_1}, ..., z_{j_l}\}$, and $z_{i_1} = x_{i_1}$. Then taking the sum for the value of x_{i_1} first, we obtain equation (10) by using lemma 2. The proof of equation (11) is similar. ∎

Lemma 4 (Decomposition of Variance)

The variance of the fitness function f can be decomposed into

$$Var_p = V_1 + V_2 + \cdots + V_{n-1} + V_n. \tag{12}$$

Here V_k for $k = 1$ to $n-1$ are defined as

$$V_k = \sum_{\substack{(i_1,...,i_k) \\ i_1<...<i_k}} \sum_{x_{i_1},...,x_{i_k}} p_{(i_1,...,i_k)}(x_{i_1}, ..., x_{i_k})\, f_{(i_i,...,i_k)}(x_{i_i}, ..., x_{i_k})^2, \tag{13}$$

and

$$V_n = \sum_{\mathbf{x}} p(\mathbf{x})\, r_{n-1}(\mathbf{x})^2. \tag{14}$$

Proof This decomposition is a direct consequence of the orthogonality described in lemma 3 ∎

The above decomposition is also used in the field of experimental design, factor analysis and analysis of variance. We are now ready to formulate the main theorem.

4 Decomposition of covariance

In order to estimate the regression coefficent, the covariance will be decomposed.

Theorem 2 *In a random mating genetic population with uniform crossover and in linkage equilibrium, the equation*

$$Cov_{\bar{p}o} = \frac{1}{2}V_1 + \frac{1}{4}V_2 + \cdots + \frac{1}{2^n}V_n = \sum_{k=1}^{n} \frac{1}{2^k}V_k \tag{15}$$

holds.

Proof We recall the definition of the midparent–offspring covariance:

$$Cov_{\bar{p}o} = \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y}) \left(\frac{f(\mathbf{x}) + f(\mathbf{y})}{2} - \bar{f} \right) \frac{1}{2^n} \sum_{i=1}^{2^n} (f(\mathbf{z}_i) - \bar{f}).$$

Here each element of $\mathbf{z}_i$ is inherited from one of the parents' alleles. Without loss of generality we can assume $\bar{f} = 0$. Because of random mating we have

$$Cov_{\bar{p}o} = \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})\ f(\mathbf{x})\ \frac{1}{2^n}(f(\mathbf{z}_1) + \cdots + f(\mathbf{z}_{2^n})).$$

We now compute one of the 2^n elements

$$\sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})f(\mathbf{x})f(\mathbf{z}_i).$$

These elements can be classified into $n+1$ classes according to the number of alleles (0 to n) which $\mathbf{z}_i$ inherits from $\mathbf{x}$.

For example, let $\mathbf{x}$ and $\mathbf{z}_i$ have one common allele at locus 1, that is, $\mathbf{z}_i = (x_1, y_2, ..., y_n)$. Then applying the decomposition of $f(\mathbf{x})$ and using the orthogonality lemma 3 repeatedly, we get the result

$$\begin{aligned} \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})f(\mathbf{x})f(\mathbf{z}_i) &= \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})(f_{(1)}(x_1) + f_{(2)}(x_2) + \cdots + r_{n-1}(\mathbf{x})) \\ &\times \ (f_{(1)}(x_1) + f_{(2)}(y_2) + \cdots + r_{n-1}(\mathbf{z}_i)) \\ &= \sum_{x_1} p_1(x_1) f_{(1)}(x_1)^2. \end{aligned}$$

Now summing over all $\mathbf{z}_i$ which inherit one allele from $\mathbf{x}$, we obtain

$$C_1 = \sum_{\mathbf{z}_i,\ |\mathbf{x}\cap\mathbf{z}_i|=1} \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})f(\mathbf{x})f(\mathbf{z_i}) = \sum_i \sum_{x_i} p_i(x_i) f_{(i)}(x_i)^2 = V_1.$$

In general, let $\mathbf{x}$ and $\mathbf{z}_i$ have s common alleles. Without loss of generality let $\mathbf{z}_i = (x_1, x_2, ..., x_s, y_{s+1}, ..., y_n)$ then

$$\begin{aligned}
\sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})f(\mathbf{x})f(\mathbf{z}_i) &= \sum_{i=1}^{s} \sum_{x_i} p_i(x_i) f_{(i)}(x_i)^2 \\
&+ \sum_{\substack{(i,j)\\ i<j}} \sum_{x_1,x_j} p_i(x_i)p_j(x_j)\ f_{(i,j)}(x_i,x_j)^2 + \cdots \\
&+ \prod_{i=1}^{s} p_i(x_i)\ f_{(1,...,s)}(x_1,...,x_s)^2.
\end{aligned}$$

The above equation is a result of the linkage equilibrium. Carefully counting the number of instances of each term and taking the sum of all possible combinations of $\mathbf{x}$ and $\mathbf{z}_i$ which have s alleles in common, we get

$$\begin{aligned}
C_s &= \sum_{\mathbf{z}_i,\ |\mathbf{x}\cap\mathbf{z}_i|=s} \sum_{\mathbf{xy}} p(\mathbf{x})p(\mathbf{y})f(\mathbf{x})f(\mathbf{z_i}) \\
&= \left(\begin{array}{c} n-1 \\ s-1 \end{array}\right) V_1 + \left(\begin{array}{c} n-2 \\ s-2 \end{array}\right) V_2 + \cdots + \left(\begin{array}{c} n-s \\ 0 \end{array}\right) V_s.
\end{aligned}$$

Finally by summing up all C_s from $s = 1$ to n, we obtain

$$Cov_{\bar{p}o} = \frac{1}{2^n} \sum_{s=1}^{n} C_s = \frac{1}{2} V_1 + \frac{1}{2^2} V_2 + \cdots + \frac{1}{2^n} V_n \quad \blacksquare$$

From this theorem, we can easily derive two corollaries.

Corollary 1

$$Cor_{\bar{p}o} = \sum_{k=1}^{n} \frac{\sqrt{2}}{2^k} \frac{V_k}{Var_p}, \qquad b_{\bar{p}o} = \sum_{k=1}^{n} \frac{1}{2^{k-1}} \frac{V_k}{Var_p}. \tag{16}$$

Corollary 2 *If the fitness function f is additive, that is, $f(\mathbf{x}) = \sum_i f_{(i)}(x_i)$, then*

$$Cor_{\bar{p}o} = \frac{1}{\sqrt{2}}, \qquad b_{\bar{p}o} = 1. \tag{17}$$

5 Examples and comparison with simulations

In [9] the above theorem was applied to some popular test functions. Numerically decomposing the variance is computationally far too expensive to be of use for the breeder genetic algorithm. Breeders conjecture that the *additive genetic variance* V_1 is the most important factor of the heritability. The higher order interactions will contribute much less to the heritability. We will test in a forthcoming paper if this conjecture is correct. Here we give some simple examples.

5.1 ONEMAX function

Let each gene have n loci, and each locus have 2 alleles 0 and 1. The fitness value is defined by the number of 1s in the gene. This function is totally additive. Hence for ONEMAX function we may use corollary 2.

$$Cor_{\bar{p}o} = \frac{1}{\sqrt{2}}, \qquad b_{\bar{p}o} = 1.$$

Results from Monte-Carlo type simulation using our Parallel Genetic Algorithm Simulator PeGAsuS confirm the above values.

5.2 PLATEAU function

Let the chromosome be composed of n building blocks and each building block have k loci. Each loci has 2 alleles 0 and 1. Then PLATEAU(n, k)(**x**) is defined by the number of blocks which contain only 1's. If this number is s then the fitness value is ks. Here we take PLATEAU(10,3) as an example. We decomposed it up to 2nd order terms under the condition of $p_i(x_i) = 1/2$. The decomposition of the variance is given by $V_p = V_1 + V_2 +$higher order term. We calculated

$$\frac{V_1}{Var_p} = \frac{V_2}{Var_p} = \frac{3}{7}.$$

The regression coefficient $b_{\bar{p}o}$ is given by

$$b_{\bar{p}o} = \frac{V_1}{Var_p} + \frac{1}{2}\,\frac{V_2}{Var_p} + \text{higher order term.}$$

The first two terms give $b_{\bar{p}o} = 0.64$. The effect of higher order terms is less than $1/28 = 0.03$. In this case PeGASuS gives the result $b_{\bar{p}o} \approx 0.65$ for PLATEAU(10,3).

5.3 DECEPTIVE function

DECEPTIVE(n,k)(**x**) is defined on nk loci each with two alleles. For the definition of this function please see [5]. Here we consider DECEPTIVE(1,3). We decomposed it up to 3rd order terms under the condition of $p_i(x_i) = 1/2$ and we got

$$\frac{V_1}{Var_p} = \frac{21}{155} \quad \frac{V_2}{Var_p} = \frac{70}{155} \quad \frac{V_3}{Var_p} = \frac{64}{155}.$$

The estimated value of regression coefficient is $b_{\bar{p}o} = 0.465$ From the result of our simulations, it is expected that the value of the regression coefficient does not depend strongly on n, the number of blocks. Simulation show a highly oscillating regression coefficient. The average is given by $b_{\bar{p}o} = 0.5$ for DECEPTIVE(10,3).

6 Conclusion

We have derived a formula for calculating the midparent–offspring regression coefficient from the microscopic genetic information of the population. This result was applied to three

popular fitness functions. The value calculated by the formula was confirmed by Monte Carlo type simulation.

In [9] Mühlenbein et.al. it is discussed how the above method is used for the Breeder Genetic Algorithm BGA. The BGA is part of the simulation environment PeGAsuS. The numerical implementation of the general decomposition method is prohibitive. It will be useful only if the first term, *the additive genetic variance* V_1 is the main factor contributing to the heritability. This is conjectured in the scientific literature about breeding.

The additive genetic variance is also used in Fisher's well known *Fundamental Theorem of Natural Selection*. It postulates that the heritability which can be exploited by selection only depends on the additive part of the genetic variance[2]. We will investigate this conjecture in the future.

We will also show the connection of our fundamental theorem to the schema theorem used in the literature about genetic algorithms [5]. Extending our result to continuous fitness functions (infinite number of alleles), noisy fitness functions and other crossover operation is also worthwhile to investigate.

Acknowledgements

Bill Buckles carefully read the manuscript and gave us helpful comments. This work was done while one of the authors (Hideki Asoh) was at GMD as a guest researcher. He thanks GMD for giving the chance of staying, and also to the Science and Technology Agency in Japan for supporting his stay. This work is a part of the SIFOGA project supported by Real World Computing Partnership.

References

[1] Bäck,T. and Schwefel,H.-P. An overview of evolutionary algorithms for parameter optimization, *Evolutionary Computation* **1**, 1-24, 1993.

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

[3] Fisher,R.A. *The Genetical Theory of Natural Selection 2nd. Ed.*, Dover Press, New York, 1958.

[4] Freedman,D., Pisani,R., Purves,R., and Adhikkari,A. *Statistics* 2nd. Ed., W.W.Norton, New York, 1991.

[5] Goldberg,D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Readin, 1989.

[6] Hartl,D.L. *A Primer of Population Genetics*, Sinauer Associates, Sunderland, 1981.

[7] Mühlenbein,H. Evolutionary algorithms: Theory and applications, in E.Aarts and J.K.Lenstra (eds.) *Local Search in Combinatorial Optimization*, Wiley, 1993.

[8] Mühlenbein,H.and Schlierkamp-Voosen,D. Predictive models for the Breeder Genetic Algorithm, *Evolutionary Computation* **1**, 25-49, 1993.

[9] Mühlenbein,H. and Schlierkamp-Voosen,D. The science of breeding and its application to the breeder genetic algorithm BGA, *Evolutionary Computation*, 1994 (to be published).