Exact analysis of selection and recombination for two and three loci

Heinz Mühlenbein[*] Dragan Cvetković[†]
GMD, 53754 St. Augustin, Germany

Abstract

Deterministic equations are derived describing the evolution of gene frequencies for two and three loci with binary alleles. Even ONEMAX function leads to nonlinear difference equations. An exact expression for the realized heritability is derived. It is shown with simple fitness functions that recombination may fail to lead the population to the optimum. The challenge of population genetics and genetic algorithm research remains to find approximate solutions to the exact equations. Our approximation for the ONEMAX function, based on the assumption of binomial distributed fitness values, reasonably predicts the evolution of the average fitness.

1 Introduction

A new approach to a predictive theory of genetic algorithms based on population genetics was introduced in [MSV93], [MSV94], [AM94]. A key concept of this approach is the equation for the *response to selection* introduced by breeders. In [MSV94] an approximate solution of the response to selection equation has been computed for the ONEMAX function with recombination and proportionate or truncation selection. The agreement of the theoretical results with simulation results was promising.

In this paper we investigate the exact equations describing the dynamic behavior of genetic populations. We restrict our analysis to two and three loci. The exact equation for the response to selection is derived. It turns out, that the exact equations seem to be analytically intractable, even for the ONEMAX function of two loci. This result, at first sight surprising, is well known in population genetics [Nag92].

[*] Heinz.Muehlenbein@gmd.de
[†] Dragan.Cvetkovic@gmd.de

Population genetics distinguishes between three levels of description of the evolution of a population, whether natural or artificial [Nag92]. The specification of the variables of interest as functions of generations (or time) is a *complete solution*. A complete solution describes the dynamic behavior of a genetic algorithm. It can be used to determine the speed of convergence for the genetic algorithm. Unfortunately a complete solution has only been obtained for a few cases. However, one can often determine the fate of the population for all initial conditions. This description is called a *complete* or *global analysis*. A global analysis can be used to decide whether a genetic algorithm will converge to the global optimum. If a complete analysis cannot be carried out, one may still obtain some information of evolutionary interest by locating all the equilibria of the dynamic system. Then the behavior of the system in the neighborhood of these stationary states can be investigated. This analysis is called a *local analysis*. A local analysis is of limited value for genetic algorithm research. Global and local analyze for some two loci systems were done in [JV94].

Unfortunately, many investigations of classical population genetic models are mathematically opaque. In order to facilitate the analysis additional assumptions are often introduced during proofs. Today the emphasis in population genetics is to reinvestigate some of the classical models. A recent survey of population genetics was done by Naglyaki [Nag92]. He writes in section 10.5: "Since a basic rigorous theory of evolutionary time scales is just being developed, we restrict ourselves to the short-term effect of selection." A new theory of selection is presented in [TB94]. There, the equation for the response to selection is investigated for an additive polygenetic trait.

In this paper the long-term effects of selection and recombination are analyzed for binary alleles. The investigation is restricted to proportionate selection.

2 The response to selection for two loci

For two loci there are four possible genotypes: $((0,0),(0,1),(1,0),(1,1))$ indexed by $i = (0,1,2,3)$. We denote their fitness values (m_0, m_1, m_2, m_3). Let $q_i(t)$ be the frequency of genotype i at generation t. We assume an infinite population and we use Mendelian recombination, also called *uniform crossover* in genetic algorithms literature [Sys89]. For proportionate selection the exact equations describing the evolution of the frequencies q_i can easily be derived. The equations are well known for diploid chromosomes in population genetics [CK70]. Therefore we just state the equations

$$q_i(t+1) = \frac{m_i}{\bar{f}(t)} q_i(t) + \epsilon_i \frac{D(t)}{2\bar{f}(t)^2} \quad i = 0,1,2,3 \tag{1}$$

where ϵ is $(-1,1,1,-1)$ and where $\bar{f}(t) = \sum_{i=0}^{3} m_i q_i(t)$ is the average fitness of the population. $D(t)$ defines the deviation from linkage equilibrium

$$D(t) = m_0 m_3 q_0(t) q_3(t) - m_1 m_2 q_1(t) q_2(t) \tag{2}$$

From equation (1) one obtains the exact equation for the response

$$R(t) = \bar{f}(t+1) - \bar{f}(t) = \frac{V(t)}{\bar{f}(t)} - (m_0 + m_3 - m_1 - m_2)\frac{D(t)}{2\bar{f}(t)^2} \tag{3}$$

where $V(t) = \sum q_i(t)(m_i - \bar{f}(t))^2$ denotes the phenotypic variance of the population. For proportionate selection the selection differential $S(t)$ and the variance are related through the following equality ([MSV94])

$$S(t) = \frac{V(t)}{\bar{f}(t)} \tag{4}$$

Combining the above equations we obtain the exact equation for the response to selection $R(t) = b(t)\,S(t)$ where

$$b(t) = 1 - (m_0 - m_1 - m_2 + m_3)\,\frac{D(t)}{2\bar{f}(t)V(t)} \tag{5}$$

$b(t)$ is called the *realized heritability* in population genetics. In order to predict $R(t)$ for a number of generations, $b(t)$ and $S(t)$ have to be estimated. In general, $b(t)$ depends on the gene frequencies of the population. For fitness functions with $(m_0 + m_3 = m_1 + m_2)$ one obtains $b(t) = 1$. Several methods for estimating $b(t)$ are discussed elsewhere [MSV94]. In this paper we concentrate on estimating $S(t)$. First we consider the ONEMAX function.

3 The ONEMAX function

The function value of ONEMAX is just the sum of the occurrences of allele 1. For ONEMAX an approximate solution of the response to selection has been computed for an arbitrary number of loci n in [MSV94]. The solution is given in terms of $p(t)$, the probability of allele 1. The equation was solved under the assumption that the variance of the fitness is binomial distributed [MSV94].

$$V(t) = n\,p(t)\,(1 - p(t)) \tag{6}$$

where n is the number of loci. We denote the solution obtained with this assumption as $p_{bin}(t)$. From $\bar{f}_{bin}(t) = n \cdot p_{bin}(t)$ and from $R(t) = V(t)/\bar{f}(t)$ one obtains $R_{bin}(t) = 1 - p_{bin}(t)$. This gives the recurrence equation

$$p_{bin}(t+1) = p_{bin}(t) + \frac{1}{n}(1 - p_{bin}(t)) \tag{7}$$

This recurrence equation has the solution

$$p_{bin}(t) = 1 - (1 - \frac{1}{n})^t (1 - p(0)) \tag{8}$$

For the average fitness $\bar{f}(t)$ the agreement with simulation results is promising [MSV94].

We will now compare the approximate equations with the exact equations for two loci. The fitness values for ONEMAX are given by $(0, 1, 1, 2)$. From (1) we obtain

$$\begin{aligned} \bar{f}(t+1) &= q_1(t+1) + q_2(t+1) + 2q_3(t+1) \\ &= \frac{q_1(t) + q_2(t) + 4q_3(t)}{\bar{f}(t)} \\ &= 1 + \frac{2q_3(t)}{\bar{f}(t)}. \end{aligned} \tag{9}$$

By definition $\bar{f}(t) = 2p(t)$. Therefore

$$R(t) = 1 - p(t) + \frac{B_3(t)}{p(t)} \tag{10}$$

where $B_3(t)$ denotes how far $q_3(t)$ deviates from the frequency given by the binomial distribution:

$$B_3(t) = q_3(t) - p^2(t) \tag{11}$$

The difference equation for $p(t)$ can be written as

$$p(t+1) = p(t) + \frac{1}{2}(1 - p(t)) + \frac{B_3(t)}{2p(t)} \tag{12}$$

The approximate equation (7) is obtained if $B_3(t) = 0$. Simulations show that this assumption is not fulfilled. How large $B_3(t)$ is, cannot be estimated from the difference equations. We have not yet succeeded to analytically solve the exact difference equations.

Table 1 shows some numerical results. The approximate analysis gives quite different results for the second and third generation. But after 10 generations the results are almost equal. In fact, we have $p(t+1) \approx p_{bin}(t)$.

We now turn to a global analysis for certain nonlinear fitness functions.

4 Equilibrium analysis for nonlinear fitness functions

In this section we analyze the equilibria for two classes of nonlinear fitness functions. The first function has two optima and is defined by

$$(c \cdot m, m, m, c \cdot m) \qquad c \geq 0$$

t	$R(t)$	$R_{bin}(t)$	$p(t)$	$p_{bin}(t)$	$f(t)$	$f_{bin}(t)$	$B_3(t)$	$q_3(t)$
0	0.875	0.875	0.125	0.125	0.25	0.25	0	0.01563
1	0.26736	0.4375	0.5625	0.5625	1.125	1.125	-0.09570	0.22070
2	0.23752	0.30382	0.69618	0.78125	1.39236	1.5625	-0.04615	0.43851
3	0.16405	0.18506	0.81494	0.89062	1.62989	1.78125	-0.01712	0.64701
4	0.09712	0.10303	0.89697	0.94531	1.79393	1.89062	-0.00531	0.79924
5	0.05291	0.05448	0.94552	0.97266	1.89105	1.94531	-0.00148	0.89253
6	0.02762	0.02802	0.97198	0.98633	1.94395	1.97266	-0.00039	0.94435
7	0.01411	0.01421	0.98579	0.99316	1.97157	1.98633	-0.00010	0.97167
8	0.00713	0.00716	0.99284	0.99658	1.98568	1.99316	-0.00003	0.98571
9	0.00358	0.00359	0.99641	0.99829	1.99282	1.99658	-0.00001	0.99282

Table 1: Results for the equations (7)–(12)

The equilibria can be obtained from

$$q_i(t+1) = q_i(t) \qquad i = 0, 1, 2, 3 \tag{13}$$

We assume that the initial gene frequencies fulfill $q_1(0) = q_2(0)$. Then we have for all t $q_1(t) = q_2(t)$. After simple but tedious manipulations one obtains

$$q_1^* = \frac{3 - 5c + \sqrt{9(1-c)^2 + 4c}}{16(1-c)} \tag{14}$$

There exists a unique equilibrium given by

$$E^* = (0.5 - q_1^*, q_1^*, q_1^*, 0.5 - q_1^*)$$

It is interesting to note that even for the case $c \rightarrow \infty$ there is a stable polymorphism. The genetic algorithm does not converge to the global optima, but to a distribution containing all genotypes. The equilibrium is given by

$$E^* = (0.375, 0.125, 0.125, 0.375).$$

For the case $c = 0$ one obtains

$$E^* = (0.125, 0.375, 0.375, 0.125).$$

The above result shows how difficult it is to predict the evolution of genetic populations. For nonlinear fitness functions recombination can be antagonistic to selection and can create genotypes of very low fitness.

Next we give an example where the genetic algorithm with proportionate selection does not converge to the global optimum, but only to a local optimum. We consider the following nonlinear fitness function

$$(1-c,0,0,1) \tag{15}$$

Using $Mathematica^{TM}$, the following three equilibrium points were obtained

$$(1,0,0,0)\,,(0,0,0,1)\,,\left(\frac{3\,(1+c)}{2\,(2-c)^2},\frac{1-c-2\,c^2}{2\,(2-c)^2},\frac{1-c-2\,c^2}{2\,(2-c)^2},\frac{3\,(1-3\,c+2\,c^2)}{2\,(2-c)^2}\right)$$

Depending on the initial gene distribution, the population will converge to one of the equilibrium points. We numerically investigated the special case $c = 0.01$. For $p(0) = 0.45$ the population converges to the genotype (00) (see Table 2). For $p(0) = 0.55$ the population converges to the global optimum (11) (see Table 3). Note how sharply the realized heritability drops after the first generation. The third equilibrium is isolated. If the initial population is already in this state, it remains there. Points in the neighborhood of the equilibrium will converge either to the first or the second equilibrium. The convergence is extremely slow.

t	$q_0(t)$	$q_1(t)$	$q_2(t)$	$q_3(t)$	$f(t)$	$R(t)$	$S(t)$	$b(t)$
0	0.3025	0.2475	0.2475	0.2025	0.5020	0.2526	0.4921	0.5133
1	0.4763	0.1203	0.1203	0.2831	0.7546	0.0059	0.2392	0.0249
2	0.5077	0.1172	0.1172	0.2579	0.7605	0.0099	0.2329	0.0424
3	0.5488	0.1121	0.1121	0.2271	0.7704	0.0157	0.2226	0.0706
4	0.6013	0.1039	0.1039	0.1908	0.7861	0.0234	0.2063	0.1135
5	0.6653	0.0919	0.0919	0.1509	0.8095	0.0315	0.1824	0.1727
6	0.7378	0.0758	0.0758	0.1105	0.8410	0.0367	0.1503	0.2443
7	0.8115	0.0571	0.0571	0.0744	0.8777	0.0360	0.1131	0.3180
8	0.8765	0.0388	0.0388	0.0459	0.9137	0.0293	0.0768	0.3813
9	0.9258	0.0239	0.0239	0.0264	0.9430	0.0202	0.0473	0.4274
10	0.9584	0.0136	0.0136	0.0144	0.9632	0.0123	0.0270	0.4566
11	0.9777	0.0074	0.0074	0.0076	0.9755	0.0069	0.0146	0.4734
12	0.9884	0.0039	0.0039	0.0039	0.9824	0.0037	0.0076	0.4826
13	0.9940	0.0020	0.0020	0.0020	0.9861	0.0019	0.0039	0.4874
14	0.9970	0.0010	0.0010	0.0010	0.9880	0.0010	0.0020	0.4899

Table 2: Results for function (15) for $c = 0.01$ and $p_0 = 0.45$.

5 The exact equations for three loci

In principle it is possible to write down the exact equations for proportionate selection and recombination for any arbitrary number of loci. But in our opinion these equations are of limited use. First, the number of genotypes increases exponentially in the number of loci. For 10 loci there are 2^{10} different genotypes. Second, the two loci case has not yet been solved. Even good approximations have not been obtained.

t	$q_0(t)$	$q_1(t)$	$q_2(t)$	$q_3(t)$	$f(t)$	$R(t)$	$S(t)$	$b(t)$
0	0.2025	0.2475	0.2475	0.3025	0.5030	0.2545	0.4930	0.5162
1	0.2787	0.1199	0.1199	0.4816	0.7575	0.0084	0.2389	0.0353
2	0.2485	0.1158	0.1158	0.5199	0.7659	0.0139	0.2309	0.0603
3	0.2122	0.1090	0.1090	0.5698	0.7799	0.0216	0.2174	0.0995
4	0.1709	0.0984	0.0984	0.6323	0.8015	0.0307	0.1964	0.1561
5	0.1279	0.0833	0.0833	0.7056	0.8322	0.0380	0.1663	0.2284
6	0.0876	0.0645	0.0645	0.7834	0.8702	0.0395	0.1288	0.3069
7	0.0548	0.0449	0.0449	0.8554	0.9097	0.0339	0.0897	0.3778
8	0.0316	0.0280	0.0280	0.9123	0.9436	0.0242	0.0561	0.4312
9	0.0171	0.0160	0.0160	0.9508	0.9678	0.0149	0.0321	0.4656
10	0.0089	0.0086	0.0086	0.9739	0.9827	0.0084	0.0172	0.4856
11	0.0045	0.0045	0.0045	0.9866	0.9911	0.0044	0.0089	0.4963
12	0.0023	0.0023	0.0023	0.9932	0.9955	0.0023	0.0045	0.5019
13	0.0011	0.0011	0.0011	0.9966	0.9977	0.0011	0.0023	0.5047
14	0.0006	0.0006	0.0006	0.9983	0.9989	0.0006	0.0011	0.5061

Table 3: Results for function (15) for $c = 0.01$ and $p_0 = 0.55$.

The general recombination equation for an arbitrary number of loci can be generated from a program written by Martin Baatz (personal communication). Just to give a feeling for how complex the equations become, here is the exact equation for $n = 3$ loci:

$$R(t) = b(t)\, S(t) \tag{16}$$

$$\begin{aligned} b(t) = {} & 1 + [(m_0 - m_1 - m_2 + m_3)\,(m_1\, m_2\, q_1\, q_2 - m_0\, m_3\, q_0\, q_3) \\ & + (m_0 - m_1 - m_4 + m_5)\,(m_1\, m_4\, q_1\, q_4 - m_0\, m_5\, q_0\, q_5) \\ & + (m_0 - m_2 - m_4 + m_6)\,(m_2\, m_4\, q_2\, q_4 - m_0\, m_6\, q_0\, q_6) \\ & + (m_1 - m_3 - m_5 + m_7)\,(m_3\, m_5\, q_3\, q_5 - m_1\, m_7\, q_1\, q_7) \\ & + (m_2 - m_3 - m_6 + m_7)\,(m_3\, m_6\, q_3\, q_6 - m_2\, m_7\, q_2\, q_7) \\ & + (m_4 - m_5 - m_6 + m_7)\,(m_5\, m_6\, q_5\, q_6 - m_4\, m_7\, q_4\, q_7)]/(2\, \bar{f}\, V) \\ & + [\sum_{i=0}^{3} m_i\, m_{7-i}\, q_i\, q_{7-i}\, (\sum_{k=0}^{7} m_k - 4\,(m_i + m_{7-i}))]/(4\, \bar{f}\, V) \end{aligned} \tag{17}$$

For fitness functions contained in the class of *unification* functions, the number of equations may be reduced to the number of loci, if the initial gene distribution is chosen appropriately. Unification means that the fitness values depend only on the number of occurrences of allele 1 in the genotype. For this class of fitness functions, one may assume that the gene frequencies for genotypes with the same number of occurrences of allele 1 are equal.

For the class of unification functions, equation (17) reduces to[1]

$$\begin{aligned} b(t) &= 1 + \frac{3(a_0 + a_2 - 2\,a_1)}{2\bar{f}V} \begin{vmatrix} a_1\,r_1 & a_0\,r_0 \\ a_2\,r_2 & a_1\,r_1 \end{vmatrix} + \frac{3(a_1 + a_3 - 2\,a_2)}{2\bar{f}V} \begin{vmatrix} a_2\,r_2 & a_1\,r_1 \\ a_3\,r_3 & a_2\,r_2 \end{vmatrix} \\ & \quad + \frac{3(a_1 + a_2 - a_0 - a_3)}{4\bar{f}V} \begin{vmatrix} a_0\,r_0 & a_1\,r_1 \\ a_2\,r_2 & a_3\,r_3 \end{vmatrix} \end{aligned} \tag{18}$$

where

$$\begin{array}{ll} m_0 = a_0 & q_0 = r_0 \\ m_1 = m_3 = m_4 = a_1 & q_1 = q_2 = q_4 = r_1 \\ m_3 = m_5 = m_6 = a_2 & q_3 = q_5 = q_6 = r_2 \\ m_7 = a_3 & q_7 = r_3 \end{array}$$

We can easily see that $b(t) = 1$ if $a_0 + a_2 = 2\,a_1$ and $a_1 + a_3 = 2\,a_2$. This assumption is fulfilled for ONEMAX, but also for $a_0 = 1 - c$, $a_1 = 1 - 2c/3$, $a_2 = 1 - c/3$ and $a_3 = 1$ (for $0 \le c \le 1$). The heritability of the latter function is 1, but the convergence can be very slow because of a small variance. Our simulations showed that the population always converges to the global optimum.

The sometimes antagonistic interaction between recombination and selection shows clearly up for three loci. We take as an example a well known deceptive problem, defined by the fitness function

$$(1-c, 1-2c, 1-2c, 0, 1-2c, 0, 0, 1) \qquad 0 \le c \le 0.5 \tag{19}$$

The genotypes are ordered according to $(000, 001, 010, 011, 100, 101, 110, 111)$. The above function belongs to the class of unification functions. Let $r_i(t)$ denote the frequency of genotypes containing i 1's in generation t. Iterating the exact difference equations produces the numerical results of Table 4.

Note that even for $p_0 = 0.5$ the population converges to the local optimum. The convergence is very slow because the realized heritability is low.

6 Conclusion

Deterministic equations were analyzed that describe the dynamics of a genetic population under proportionate selection and recombination. The equations describe the change of the gene frequencies after selection and mating. The exact difference equations for two loci could not be solved, even in the case of the ONEMAX fitness function. We have numerically shown that our approximate solution is very accurate for long term prediction of the average fitness. Some preliminary results have been presented for three loci.

We have serious doubts that the exact equations can be solved analytically, even for a small number of loci. Progress can only be expected if statistical

[1] The expression within the vertical bars denote the determinant.

t	$r_0(t)$	$r_1(t)$	$r_2(t)$	$r_3(t)$	$f(t)$	$R(t)$	$S(t)$	$b(t)$
0	0.1250	0.3750	0.3750	0.1250	0.6162	0.1617	0.3698	0.4372
1	0.2295	0.4479	0.2108	0.1118	0.7780	0.0443	0.2078	0.2131
2	0.3137	0.4490	0.1656	0.0717	0.8223	0.0458	0.1633	0.2804
3	0.4007	0.4411	0.1191	0.0392	0.8681	0.0395	0.1174	0.3364
4	0.4828	0.4195	0.0792	0.0184	0.9075	0.0278	0.0781	0.3559
5	0.5529	0.3880	0.0514	0.0077	0.9353	0.0171	0.0507	0.3372
6	0.6092	0.3534	0.0345	0.0029	0.9523	0.0101	0.0340	0.2970
7	0.6538	0.3205	0.0246	0.0011	0.9624	0.0062	0.0243	0.2551
8	0.6898	0.2912	0.0187	0.0004	0.9686	0.0041	0.0184	0.2228
9	0.7192	0.2658	0.0148	0.0001	0.9727	0.0029	0.0146	0.1986
10	0.7439	0.2440	0.0121	0.0000	0.9756	0.0022	0.0119	0.1835
11	0.7649	0.2251	0.0100	0.0000	0.9778	0.0017	0.0099	0.1718

Table 4: Results for function (19) for $c = 0.01$ and $p_0 = 0.5$

assumptions concerning the distribution of the variance can be made. Steps in this direction have been made for discrete fitness functions [AM94] and for continuous fitness functions [TB94]. For continuous functions, variance estimates can be obtained more easily. Here progress seems to be more likely. In fact, the new recombination operators of the Breeder Genetic Algorithm for continuous function optimization have been analyzed according to their variance reduction.

Now time has come for a closer cooperation between population-genetics research and genetic-algorithm research. The challenge is to find approximative solutions of the exact equations which reasonably predict the evolution of the average fitness. Genetic algorithm can supply all the data population genetics dreamed off. The theory presented here was successfully used for calibrating the implementation of our breeder genetic algorithm. Furthermore, the theory has been used to design new recombination operators and to get rid off a number of parameters controlling the algorithm.

References

[AM94] H. Asoh and H. Mühlenbein. Estimating the heritability by decomposing the genetic variance. In Y. Davidor, H.-P. Schwefel, and R. Männer, editors, *Parallel Problem Solving from Nature*, Lecture Notes in Computer Science 866, pages 98–107. Springer-Verlag, 1994.

[CK70] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Harper and Row, New York, 1970.

[JV94] J. Juliany and M. D. Vose. The genetic algorithm fractal. *Evolutionary Computation*, 2:165–180, 1994.

[MSV93] H. Mühlenbein and D. Schlierkamp-Voosen. Predictive Models for the Breeder Genetic Algorithm I. Continuous Parameter Optimization. *Evolutionary Computation*, 1:25–49, 1993.

[MSV94] H. Mühlenbein and D. Schlierkamp-Voosen. The science of breeding and its application to the breeder genetic algorithm. *Evolutionary Computation*, 1:335–360, 1994.

[Nag92] T. Naglyaki. *Introduction to Theoretical Population Genetics*. Springer, Berlin, 1992.

[Sys89] G. Syswerda. Uniform crossover in genetic algorithms. In H. Schaffer, editor, *3rd Int. Conf. on Genetic Algorithms*, pages 2–9, San Mateo, 1989. Morgan Kaufmann.

[TB94] M. Turelli and N.H. Barton. Genetic and statistical analyses of strong selection on polygenic traits: What, me normal. *Genetics*, 138:913–941, 1994.