Towards an Optimal Mutation Probability for Genetic Algorithms

Jürgen Hesser bd0@dhdurz1.bitnet

Reinhard Männer maenner@dhdphy5.bitnet

Physics Institute, University of Heidelberg
Philosophenweg 12, D-6900 Heidelberg, Germany

Abstract

In this paper the optimal parameter setting of Genetic Algorithms (GAs) is investigated. Particular attention has been paid to the dependence of the mutation probability P_M upon two parameters, the dimension of the configuration space l and the population size M. Assuming strict conditions on both the problem to be optimized and the GA, P_M converges to 0 as the population size M or the dimension of the configuration space l converges to infinity. For direct application a heuristic comprising these results is presented. The parameter settings obtained by applying this heuristic are in accordance with those which have been obtained earlier by experiment.

1 Introduction

Genetic Algorithms (GAs) are parallel optimization algorithms similar to adaptation in nature. To maximize their efficiency, the three inherent parameters of GAs are to be optimized, the mutation probability P_M, the crossover probability P_C, and the population size M. For parameter optimization of GAs several results have been obtained over the last few years. De Jong and Schuster proposed heuristics for an optimal setting of the mutation probability [1][2][3], Fogarty and Booker investigated time dependencies of the mutation and the crossover probability respectively [4][5], Greffenstette and Schaffer found optimal settings for all three parameters of the GA by experiment [6][7], Goldberg and Ros estimated optimal population sizes theoretically [8][9].

This paper presents theorems that state parameter dependencies of GAs and confirm those found earlier by experiment [7] (the GA used here is described in [14]). We mostly restrict our interest to P_M since theoretical results for optimal M are available [8][9] and since results of Schaffer on a test suite of problems indicate that P_C is not a sensitive parameter [7]. Particularly the dependence of the mutation probability upon both the population size and the dimension of the configuration space is investigated using the theory of GAs [10][11], of stochastic processes [12], and of theoretical biology [13]. In addition, a heuristic is presented which exploits these results. The underlying performance measure is the online performance, i.e., the mean function value is summed up over all generations of the convergence of the GA.

2 Dependence of the Mutation Probability upon Population Size

Due to results of Holland mutation is only necessary for finite populations [10]. In finite populations genetic drift occurs, i.e., some coordinate values in the population die out due to stochastic effects of selection. Without mutation new coordinate values cannot be introduced. Therefore only a part of the configuration space can be attained during search and mostly the GA will not find the optimal solution. As a consequence, for finite populations the mutation probability should be larger than zero and it should be zero for infinite populations. This apparently necessary result is proved below.

The following definitions deal with the function for which we show the dependence described above and it deals with the permutation invariance which is needed for the assumptions of the following theorems.

2.1. DEFINITION: *The function $OneMax(x)$, which is a function of the individuals x of the population, is the sum of the coordinate values of x (see [15]).*◇

2.2. DEFINITION: *An operator $\mathcal{O}$ operating on an individual x is called permutation invariant under a permutation $\mathcal{P}$ permutating the coordinates of x if $\mathcal{P}(\mathcal{O}(x)) = \mathcal{O}(\mathcal{P}(x))$.*◇

2.3. THEOREM: *Assume an infinite population size, the function to be optimized as $OneMax$, and a crossover which is invariant under permutation of the coordinates (see definition). Under these assumptions the GA can be described as a stochastic process with $l+1$ different states.*◇

The theorem explains that the number of different states can be reduced due to permutation invariance if stochastic effects are neglected which occur only in finite populations. Because of the limited space we outline only the idea of the proof:

According to its definition, the GA can be described as a stochastic process. A state of the process is represented by a possible population. The operators of the GA change the probabilities of occurence of the different states. The number of different states is of order 2^{Ml}. When increasing the population size the stochastic effects become negligible and we can describe the GA by a deterministic process. This process is equivalent to that already described in [16] and all elements of the configuration space can be considered as the different states of the process. If the crossover operator and the function f are invariant under permutations $\mathcal{P}$ of coordinates (as guaranteed by the assumptions of theorem 2.3.) then all individuals which can be obtained by permutation of the coordinates are treated equally by the operators of the GA. Consequently, individuals are different if they differ in the number of coordinate values "1" in their chromosome. Therefore there are exactly $l+1$ different states of individuals which differ only in the function value of the individuals of that class.

2.4. THEOREM: *For the process of theorem 2.3. the optimal setting of the mutation probability is zero, $P_M = 0$, and the optimal setting of the crossover probability is one, $P_C = 1$.*$\diamond$

The reason is that mutation destroys already accumulated information and thus reduces the convergence speed. With an increasing population size the approximation of the behavior of the GA by the stochastic process of theorem 2.3. gets monotonously better. With a tighter approximation of the optimal parameter setting, the GA and the stochastic process of theorem 2.3. become increasingly similar. These results are associated with the experimental dependence $P_M \sim M^{-0.9318}$ found by Schaffer which is about $P_M \sim 1/M$ [7].

In the following part we describe the idea of the proof in more detail.

The process of theorem 2.3. starts with a uniform distribution of the coordinate values. Therefore the function values denoted by y are distributed binomially. This distribution will be denoted as $u(y)$. In the following lemma we show that we can assume at all times a distribution of the function values of the individuals x, $f(x) = y$, according to $i(y)u(y)$ where $i(y)$ is a monotonously increasing function of y (note monotonously increasing means $i(y) \leq i(y + \triangle), \triangle > 0$).

2.5. LEMMA: *Assume the crossover probability $P_C = 0$ and a starting distribution $u(y)$ (binomial distribution). Then at all times T the distribution of the function values y of the individuals x in the population can be written as $i(y)u(y)$ with i as a monotonously increasing function of y.*$\diamond$

Idea of the proof:

The lemma is proved by induction over the time T:

Consider time $T = 0$: The initial distribution of the function values y is $u(y)$ due to definition [14]. Therefore $i(y) = 1$ and $i(y)$ is monotonously increasing.

Assume the distribution of y can be written as $i(y)u(y)$ at time T. We have to show that neither selection nor mutation lead to a different distribution for the time $T + 1$.

- Selection: $i(y)u(y) \rightarrow i(y)u(y) \cdot y \cdot const.$ according to the definition of the selection [10]. But if $i(y)$ is monotonously increasing with y then $y \cdot i(y)$ is monotonously increasing with y too. So the monotony property is not altered by selection.

- Mutation: Let us consider any coordinate value. Mutation changes this value according to the mutation probability P_M. This is equivalent to replacing with probability $2P_M$ the considered coordinate value by a value "0" or "1" randomly chosen. Therefore by applying mutation often enough we obtain a random distribution of coordinate values in the population. Similarly the distribution converges to the uniform distribution $u(y)$. This convergence can be shown not to alter the property of monotony.

Since neither selection nor mutation alter the monotony property for $i(y)$, this relation is true at time $T + 1$ and therefore the lemma is proved.

To show that mutation decreases the mean function value we have to consider that the more often the function value "1" occurs in the population (i.e. the larger the mean function value $\langle f\rangle$) the more probable is this value replaced by "0" due to mutation. This signifies that the mean function value is reduced by mutation and therefore we can conclude $P_M = 0$.

For the application of the result $P_M = 0$ for all crossover probabilities, we have only to show that this result is true for nonzero P_C.

By definition the crossover operator exchanges coordinate values and therefore does not alter their frequency of occurrence. Consequently the sum of coordinate values "1" in the population (the sum is proportional to $\langle f\rangle$ due to the chosen function) is not changed by crossover, i.e., $\langle f\rangle$ is invariant under crossover. However the random exchange operation of the crossover leads to an increase of disorder in the population and therefore the variance of the function values and consequently ($\langle f\rangle$ is constant) $\langle f^2\rangle$ increases ($\langle f^2\rangle$ is the mean of the squared function values in the population). Using the result of theoretical biology, that the the mean function value after proportionate selection is $\langle f^2\rangle/\langle f\rangle$ [13], we can conclude that for a maximal $\langle f^2\rangle/\langle f\rangle$ the crossover probability P_C is optimal, i.e., $P_C = 1$. Furthermore, it can be shown that the distribution of the function values after crossover can be written as $i(y)u(y)$, with $i(y)$ as a monotonously increasing function of the function value y. For those distributions it was shown before that $P_M = 0$. Therefore this result is valid for all values of P_C.

3 Dependence of P_M upon l

The following theorem shows the dependence of the mutation probability upon the dimension of the configuration space. The theorem can be proved if for an optimal mutation probability P_M the following condition can be assumed. P_M is sufficiently small so that the mutation does not alter a fixed part kl ($k\epsilon[0,1]$) of all coordinate values for more than ε individuals, where ε is constant. The plausibilty of this assumption is described at the end of this section.

3.1. THEOREM: *Assume the population size M equal to $2^{g(l)}$, with g as any function such that $g(l)/l$ vanishes for a large dimension l of the configuration space, $g(l)/l \to 0$ for $l \to \infty$. Additionally assume the following condition for an optimal parameter setting of P_M. For more than ε individuals a fixed part kl of the coordinate values is not modified by mutation. Then P_M converges to 0 as the dimension of the configuration space l increases.*◇

This is in accordance with the previous experimental result $P_M \sim l^{-0.4535}$ [7].

Proof:

To meet the assumption of the theorem the probability not to change kl coordinate values by mutation should be larger than ε/M, i.e., on the average this combination of coordinate values is not destroyed for ε individuals. The probability for no mutation in that combination is $(1-P_M)^{kl}$. Therefore we obtain the following formula for a population size $M = 2^{g(l)}$:

$$\Rightarrow 1 - P_M > \sqrt[kl]{\frac{\varepsilon}{M}}.$$

$$\Rightarrow P_M < 1 - \sqrt[kl]{\frac{\varepsilon}{M}} = 1 - \delta 2^{-g(l)/kl} \qquad (1)$$

with $\delta = \sqrt[kl]{\varepsilon}$.

With increasing l and fixed ε, the product $\delta \cdot 2^{-g(l)/(kl)}$ converges to 1 and so the right hand side of (1) converges to 0. In other words, with larger l the probability increases to destroy combinations of good subsolutions. Accordingly, P_M has to be reduced to retain a certain survival probability for these structures.◇

The argument used in this proof is similar to that of De Jong when he suggested an optimal $P_M = 1/l$ [1]. The first assumption used in the theorem 3.1. is a population size which does not increase exponentially with the dimension of the configuration space. This condition is mostly fulfilled and obviously necessary for many practical implementations. The second assumption is that the information already accumulated in the population should not be destroyed by mutation. At the beginning of the convergence it cannot be assumed that this condition is true because only few information was accumulated. However during the convergence of the GA more and more information is stored in the population (see [17]). With a large mutation probability we cannot retain the information as was shown in the theorem. Therefore it seems reasonable that this assumption is true if the individuals are near to the optimal solution in the sense of Hamming distance.

4 Explanation of the optimal Setting of the Mutation Probability

In this section we discuss the optimal setting of P_M. With the stochastic process of theorem 2.3. we can derive the optimal parameter setting under the condition that the population size is sufficiently large. For smaller M the genetic drift [13] cannot be neglected. This drift leads to states where for some coordinates all individuals in the population have the same value. In the following this situation will be denoted as absorption. Accordingly the respective coordinate value which occurs in every individual is called absorption value.

The parameter setting is optimal if the average time to find the optimal solution T^* is minimal. This time increases when absorption occurs and when the mutation probability is greater than zero, $P_M > 0$ (see theorem 2.4.). Let T_l be the supplementary time due to mutation losses and T_w the supplementary time due to absorption (waiting time). T_l and T_w can be estimated as follows.

T_w: When absorption occurs in the population, i.e., if for some coordinates the respective values for all individuals are the same, mutation reduces the time this absorption lasts. This time is inversely proportional to the probability to escape absorption. It is the probability of a change by mutation of at least one value of the considered coordinate, i.e., $P_M \cdot M$. Thus the mean waiting time T for that change is inversely proportional to that probability of a change $T \approx 1/P_M \cdot M$. The mean waiting time contributes to the additional time T_w only if absorption occurs and if mutation is necessary. Therefore T_w is proportional to the product of the probability of absorption P_A, the probability that the absorbed value is not element of the optimal solution P_S, and the mean waiting time T, i.e., $T_w = T \cdot P_A \cdot P_S$.

Discussion of P_A: We demand that the parameters are chosen such that the optimal solution is reached with a certain fixed probability. This probability depends on the stability of the solution against destructions by mutation or against stochastic effects of selection. The loss caused by mutation is neglected because of the small mutation probability expected. Then the loss caused by selection depends on the probability of absorption P_A. In order to have constant probability to reach the optimal solution we demand that P_A is constant, i.e., particularly independent of P_M as well as independent of M and l.

Discussion of P_S: Recall, P_S is the probability that the absorbed coordinate value is not element of the optimal solution. This probability can be estimated using a special implementation of the GA [17]. According to that result, $P_S \approx \alpha' e^{-\lambda T} (\alpha' = const.)$.

Summarized $T_w = \alpha e^{-\lambda T} / P_M M (\alpha = const.)$.

Additional time due to mutation T_l: For some arguments which cannot presented due to limited space we assume the following dependencies of T_l.

- proportional to the population size M as mutation generally reduces the function value of each individual and

- proportional to the dimension of the configuration space because the reduction of the increase of the function values per generation, which is caused by mutation, depends on the "size of the problem" l.

Therefore in first order approximation the additional time due to mutation is proportional to the mutation probability, the population size, and the dimension of the configuration space, $T_l \approx \beta P_M M l, \beta = const.$

T^* can be written as $T^*(\text{absorption}, P_M > 0) = T^*(\text{no absorption}, P_M = 0)(1 + T_w + T_l)$. The optimal mutation probability can then be derived as follows:

$\frac{d}{dP_M} T^*(\text{absorption}, P_M > 0) = \frac{d}{dP_M}(const \cdot (1 + \alpha e^{-\lambda T} \frac{1}{P_M M} + \beta M l P_M)) = 0$. From that follows:

$$P_M = \sqrt{\frac{\alpha}{\beta} \frac{e^{-\lambda T/2}}{M\sqrt{l}}}. \tag{2}$$

This dependence has to be compared with the result of Schaffer [7] which is approximately

$$P_M = \frac{1.76}{M\sqrt{l}}.$$

The time dependence of P_M in (2) explains the result of Fogarty who found that reducing P_M exponentially with time increases the performance of the GA [4].

REMARK 1: For this explanation it is necessary that the population size is sufficiently large such that the optimal solution will nearly always be found. How fast the mutation probability is reduced depends on the information which is available on the problem or which can be determined during convergence.

REMARK 2: This approach also allows to determine a good mutation probability for each of the coordinates of the individuals, i.e., it allows a generalization of the mutation operator for a better adaption to the search process (similar to that of evolution strategies).

5 Heuristic Formula for Optimal Setting of P_M and P_C

For direct application, these results have been comprised into a heuristic. It originates from an approach of Eigen and Schuster [2][3] who determined an upper bound for an optimal setting of P_M for a GA without crossover which can be used as a heuristic of optimal P_M and P_C setting. The underlying idea of this heuristic is that mutation must not destroy structures (which are combinations of special coordinate values) more frequently than selection can reproduce them. Otherwise an optimal solution would not be stable in the population, i.e., the probability that the population looses these structures is not negligible.

We extended this idea to the GA with crossover and obtained a heuristic formula for the optimal setting of P_M and P_C. The probability that crossover destroys a structure is approximately [10]

$$P_C \cdot \frac{\delta(H)}{l-1} \cdot (1 - h(H)) \cdot P_{correction},$$

where $\delta(H)$ is a parameter defined in the GA theory, $h(H)$ is the frequency of structure H in the population, and $P_{correction}$ measures the non-similarity of individuals in the population according to the Hamming distance ($P_{correction} = 0$ if all individuals are identical). The respective probability for mutation is $1 - (1 - P_M)^{o(H)}$, whereby $o(H)$ is defined in [10].

The reproduction factor of selection, i.e., the number of replicants of a surviving individual (which are assumed to have a good substructure H) is denoted by S. According to the heuristic to destroy less good structures H than selection reproduces, the average number of destroyed replicants should be at most $S-1$ to guarantee that at least one replicant is not destroyed on average. The probability of destruction should therefore be at most the probability to destroy $S-1$ of S replicants, $(S-1)/S$. The probability of the destruction of a structure H can also be obtained by the respective destruction probabilities of crossover and mutation. This probability that any of these two operators destroys the structure H is $1 - (1 - P_C \frac{\delta(H)}{(l-1)} P_{correction}(1 - h(H))(1 - P_M)^{o(H)}$. Therefore we obtain the following formula

$$\frac{S-1}{S} = 1 - (1 - P_M)^{o(H)} \cdot (1 - P_C \cdot \frac{\delta(H)}{l-1} \cdot P_{correction} \cdot (1 - h(H)). \qquad (4)$$

Simplified the formula of the heuristic reads

$$\frac{S-1}{S} = o(H) \cdot P_M + P_C \cdot \frac{\delta(H)}{l-1} \cdot P_{correction} \cdot (1 - h(H)).$$

A refinement of the number of replicants with respect to M can be obtained by multiplying the left hand side of equ. (4) by $(1 - 1/\sqrt{M})$.

If P_C has to be determined P_M can be chosen according to (3).

Using our heuristic, we find that P_M increases if P_C decreases and vice versa which explains such an observation of Greffenstette [6]. When converging, the individuals of the population get more and more similar, i.e., the factors $(1 - h(H))$ and $P_{correction}$ become smaller. From formula (2) we can assume that the mutation probability does not increase with time. Therefore if the number of replicants S is constant then the crossover probability has to be increased with time. The need to increase the crossover probability when the individuals of the population get more and more similar has also been shown by Booker [5].

If we assign reasonable values for the variables in (4), we find nearly the same optimal parameter values as earlier Greffenstette and Schaffer by experiment [6][7] (Tab. 1). So we think that our heuristic has a broad application area and allows to set the parameters properly for problems which are too complicated to do this by experiments.

For a direct application of the heuristic the respective parameters S, $o(H)$, $\delta(H)$, and $P_{correction}$ have to be estimated. In most of our experiments we obtained for S a value of 2 (this value can in principle be obtained from a small test run or can be determined during convergence). For $o(H)$ and $\delta(H)$ the estimations are much more complex. One overestimation of the destruction probability is to suppose these parameters to be maximal, i.e., to have the value l. Similarly we can neglect similarities in large populations and therefore estimate $P_{correction} = 1$ and $h(H) \approx 0$. A better and time dependent estimation for the similarity measure $P_{correction}$ would consider how many coordinates have the same value for nearly all individuals in the population. Let U be the number of coordinates for which the value is uniform for 90% of the population. Then $P_{correction} \approx (l-U)/l$. One of the two remaining parameters P_M or P_C has to be determined by a different heuristic to find a result for the other parameter. For a mutation probability of 0.01 as recommended by Schaffer and Greffenstette [7][6] we obtain a crossover probability of 0.32 (see table 1).

Table 1: Optimal parameter probabilities for on-line performance (S=2 as mostly fulfilled, $o(H) = \delta(H) = l = 30$, $P_{correction} = 1$, $h(H) = 0$, $P_M = 0.01$)

	P_M	P_C
own heuristic	(0.01)	0.32
Greffenstette	0.01	0.95
Schaffer	0.02-0.002	0.95-0.25

We have applied this heuristic successfully to the optimization of Steiner Trees by GAs [14].

6 Conclusions

In this paper we showed the dependence of the mutation probability upon both the population size and the dimension of the configuration space. Interestingly the result of Schaffer can easily be explained by the estimation of the effects of mutation on the additional time to reach the solution. Additionally we obtain also the result of Fogarty that the mutation probability should be decreased during convergence. The assumptions made for this explanation regarding the frequency of absorption and the losses due to mutation can in principle be verified experimentally. This approach to determine the optimal mutation probability seems fundamental and an experimental verification of the assumptions used here is planned.

References

[1] K. De Jong. Analysis of the Behavior of a Class of Genetic Adaptive Systems. PhD. Diss., Univ. of Michigan, 1975

[2] M. Nowack; P. Schuster. Error Thresholds of Replication in Finite Populations Mutations Frequencies and the Onset of Muller's Ratchet. *J. Theor. Biol.*, **Vol. 137**, pages 375-395, 1989

[3] P. Schuster. Effects of Finite Population Size and Other Stochastic Phenomena in Molecular Evolution. *Complex Systems — Operational Approaches in Neurobiology, Physics, and Computers*, Springer, Heidelberg, 1985

[4] T.C. Fogarty. Varying the probability of mutation in the Genetic Algorithm. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 104-109, 1989

[5] L. Booker. Improving Search in Genetic Algorithms. L. Davis, *Genetic Algorithms and Simulated Annealing*, Pitman, London, 1987

[6] J.J. Greffenstette. Optimisation of Control Parameters for Genetic Algorithms. *IEEE Transactions on Systems Man and Cybernetics*, **Vol. SMC-16**, No. 1, pages 122-128, 1986

[7] J.D. Schaffer; R.A. Caruna; L.J. Eshelman; R. Das. A Study of Control Parameters Affecting Online Performance of Genetic Algorithms for Function Optimization. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 51-60, 1989

[8] D. Goldberg. Sizing Populations for Serial and Parallel Genetic Algorithms. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 70-79, 1989

[9] H. Ros. Some Results on Boolean Concept Learning by Genetic Algorithms. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 28-33, 1989

[10] J. Holland. *Adaption in Natural and Artificial Systems.* Ann Arbor, The University of Michigan Press, 1975

[11] D. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning.* Addison Wesley, Reading, MA, 1989

[12] K.L. Chung. *Markov Chains With Stationary Transition Probabilities.* Springer, Berlin, 1967

[13] J.F. Crow; M. Kimura. *An Introduction to Population Genetics Theory.* Harper & Row, New York, NY, 1970

[14] J. Hesser; R. Männer; O. Stucky. Optimization of Steiner Trees using Genetic Algorithms. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 231-236, 1989

[15] G. Syswerda. Uniform Crossover in Genetic Algorithms. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 2-9, 1989

[16] C.L. Bridges; D. Goldberg. An Analysis of Reproduction and Crossover in a Binary–Coded Genetic Algorithm. J.D. Schaffer, Proc. 3rd Int'l Conf. Genetic Algorithms & Appl., Arlington, VA, pages 28-33, 1989

[17] J. Hesser; R. Männer. An Alternative Genetic Algorithm. In these proceedings