

Noise, Fitness Distribution, and Selection Intensity in Genetic Algorithms*

Timothy Meekhof
University of Idaho
Moscow, Idaho, USA
timothym@cs.uidaho.edu

Terence Soule
University of Idaho
Moscow, Idaho, USA
tsoule@uidaho.edu

ABSTRACT

Many Genetic Algorithm (GA) problems have noisy fitness functions. In this paper, we describe a mathematical model of the noise distribution after selection and then show how this model of the noise distribution can be used to model the real, underlying selection intensity of the GA population, which promises to give us a better way to model GA convergence in the presence of noise.

Categories and Subject Descriptors

I.6 [Computing Methodologies]: Simulation and Modeling, Statistical

General Terms

Algorithms, Theory

1. INTRODUCTION

Many Genetic Algorithm (GA) problems have fitness functions with significant noise, and a body of research has been built that shows that noise strongly affects the behavior of a GAs. In particular, it has been empirically shown that in the presence of too much noise, GA search becomes essentially random search and ceases to be productive. Also, it has been shown that many design parameters such as population size, generational versus steady state, and elitism must be chosen differently in the presence of noise. [1] and [2].

Given a noisy population of a GA and a selection cut-off point, we give a model that allows us to estimate the real (without noise) distribution of the selected population. This distribution has the properties that we would expect—when noise is very large, it is essentially identical to the unselected population. But, when noise is very small, the distribution of the selected population forms a noticeable **selection differential** and a corresponding **selection intensity**.

*This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

These two values have been shown in the past to have strong and predictable relationships to the performance of a GA. Our model promises a better mechanism for understanding and predicting the convergence behavior of GAs in the presence of noise.

2. FITNESS AFTER SELECTION

We assume that the fitness of an instance *alpha* of a GA population has both a noise component and a real fitness component. We make the assumption that both the noise and the real fitness values can be modeled as drawing from (different) normal distributions. So, our fitness is:

$$\hat{F}(\alpha) = F(\alpha) + \mathcal{X} \quad (1)$$

$F(\alpha)$ is the real, or underlying fitness of population member α , and X is a random variable representing the noise. $\hat{F}(\alpha)$ is the combined noise that is visible to the GA.

Let $\Omega_F(x)$ and $\Omega_{\mathcal{X}}(x)$ denote the probability distributions of the real fitness and the noise, respectively. We make the assumption that both $\Omega_F(x)$ and $\Omega_{\mathcal{X}}(x)$ are normal.

Since $\Omega_F(x)$ is a normal distribution, we can write:

$$\Omega_F(x) = \frac{1}{\sqrt{2\pi\sigma_F^2}} e^{-\frac{(x-M_F)^2}{2\sigma_F^2}} \quad (2)$$

Let $\Omega_{F:\text{sel}}(x)$ and $\Omega_{\mathcal{X}:\text{sel}}(x)$ denote the distributions after selection. $M_{F:\text{sel}}$ is the average of the real fitness distribution after selection and M_F is the real fitness distribution before selection. σ_F^2 and $\sigma_{\mathcal{X}}^2$ are the variances of real fitness and noise, respectively. Because we are working with addition of probability distributions:

$$\Omega_{\hat{F}}(x) = \frac{\Omega_F(x) + \Omega_{\mathcal{X}}(x)}{2} \quad (3)$$

In terms of the prior distributions, the distribution of the real fitness after selection is:

$$\Omega_{F:\text{sel}}(x) = \frac{\Omega_F(x) \int_{c-x}^{\infty} \Omega_{\mathcal{X}}(t) dt}{\mathcal{Z}} \quad (4)$$

where \mathcal{Z} is a normalizing constant and c is the selection cut point, the fitness value of the worst member of the selected population.

$$\mathcal{Z} = \int_{q=-\infty}^{q=\infty} [\Omega_F(q) \int_{c-q}^{\infty} \Omega_{\mathcal{X}}(t) dt] dq \quad (5)$$

Thus, the expected mean real fitness (without noise) of the selected population is:

$$M_{F:\text{Sel}} = \int_{-\infty}^{\infty} x \Omega_{F:\text{sel}}(x) dx \quad (6)$$

Using Equation 6 to model $M_{F:\text{Sel}}$ requires that we know the average fitness of our population after before and after selection, which is usually trivial. Additionally, we need estimates for the variance of the noise, $\sigma_{\mathcal{X}}$, and the variance of the real underlying fitness, σ_F . We note that it is possible to estimate these last two using representative oversampling of some of the members of the population and then using the fact that the variances form a pythagorean triplet:

$$\sigma_F^2 = \sigma_F^2 + \sigma_{\mathcal{X}}^2 \quad (7)$$

2.1 Selection Differential and Intensity

Haldane observed in 1932 that the important metrics for analyzing fitness improvement in a breeding program are **selection differential**, S , and **selection intensity**, I . [3] Mühlenbein demonstrated their applicability to GAs in [4].

Selection differential is the difference between the mean fitness of the population before and after selection:

$$S = M_{F:\text{sel}} - M_F \quad (8)$$

Selection intensity is the ratio of selection differential and the standard deviation of the distribution:

$$I = \frac{S}{\sigma_F} \quad (9)$$

What Haldane noted in breeding programs is that the rate of change of fitness between generations in a breeding program was linearly dependent on I . Haldane also noted that the rate of change was dependent on another factor, which he called **heritability**, which captured the inherent difficulty of improving fitness over the existing population.

Heritability is not constant over the entire run of a GA, since it is dependent on position of a population in the fitness landscape. However, for a given fitness landscape and a given population, Haldane showed that the rate of improvement was linear on I for animal breeding programs, and research since then has indicated that this is true for GAs as well. [4].

Using Equation 6 to compute the average fitness of the selected population, we now have a way to compute the underlying selection intensity without noise.

In Table 2.1, we give some examples of computing selection differential using Equation 6 for various cut points and amounts of noise. Notice that the selection differential, and correspondingly the selection intensity vary tremendously based on the amount of noise.

Our technique promises to give us a better way to model techniques such as dynamic oversampling. Since we can model the real selection intensity, we should be able to choose

	$\sigma_{\mathcal{X}}$				
	0.1	0.5	1	2	10
$c = -0.5$	0.51	0.48	0.42	0.30	0.076
$c = 0$	0.79	0.71	0.56	0.36	0.080
$c = 0.5$	1.1	0.99	0.73	0.42	0.083

Table 1: This table gives the selection differential for various levels of noise $\sigma_{\mathcal{X}}$ and cut values c . It is assumed that $\sigma_F = 1$, therefore $I = S$. $S = M_{F:\text{sel}} - M_F$. Larger noise decreases the selection differential, reducing the rate of fitness improvement.

oversampling techniques that effectively balance selection intensity and evaluation costs.

3. CONCLUSION

In this paper, we have described a model for estimating the real (noiseless) fitness distribution of a GA population after selection. Even though the initial distributions are assumed to be normal, the selection process gives it a much different shape, for which our Equation 6 gives us a model.

The average of the selected fitness minus the original fitness of the population is the selection differential, and is the value that makes a GA different from a random walk. Our model gives us a way to describe this differential in the presence of noise.

4. REFERENCES

- [1] H.-G. Beyer. Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice. *Computer Methods in Applied Mechanics and Engineering*, 239-267, 2000.
- [2] Y. Jin and J. Branke. Evolutionary Optimization in Uncertain Environments—A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, 2005.
- [3] J. B. S. Haldane. *The Causes of Evolution*. Longmans, New York, 1932.
- [4] H. Mühlenbein. The equation for response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346, 1997.