# Optimal Sampling in a Noisy Genetic Algorithm for Risk-Based Remediation Design

*Gayathri Gopalakrishnan and Barbara Minsker*

Civil and Environmental Engineering, University of Illinois, Urbana, Illinois
4146 NCEL, MC-250, 205 N. Mathews Ave., Urbana, IL 61801

*David E. Goldberg*

General Engineering, University of Illinois, Urbana, Illinois
117 Transportation, MC-238, 104 S. Mathews Ave., Urbana, IL 61801

## ABSTRACT

A management model has been developed that predicts human health risk and uses a noisy genetic algorithm to identify promising risk-based corrective action designs [Smalley et al, 2000]. Noisy genetic algorithms are ordinary genetic algorithms that operate in noisy environments. The "noise" can be defined as any factor that hinders the accurate evaluation of the fitness of a given trial design. The noisy genetic algorithm uses a type of noisy fitness function called the sampling fitness function, which utilizes sampling in order to reduce the amount of noise from fitness evaluations in noisy environments. This Monte-Carlo-type sampling provides a more realistic estimate of the fitness as the design is exposed to a wide variety of conditions. Unlike Monte Carlo simulation modeling, however, the noisy genetic algorithm is highly efficient and can identify robust designs with only a few samples per design. For complex water resources and environmental engineering design problems with complex fitness functions, however, it is important that the sampling be as efficient as possible. In this paper, methods for reducing the computational effort through improved sampling techniques are investigated. A number of different sampling approaches will be presented and their performance compared using a case study of a risk-based corrective action design.

## INTRODUCTION

Simple genetic algorithms (GAs) have been used in numerous engineering design applications within water resources [Wang 1991, Ritzel et al, 1994] to identify optimal solutions. One of the principal reasons for using a GA as compared to more traditional optimization methods is that the decision space is searched from an entire population of possible designs. This allows the GA to solve discrete, non-convex, discontinuous problems without differentiation [Goldberg, 1989]. However, using the GA for real world problems can present some difficulties. One of these is the presence of "noise" in the system. The "noise" which exists in certain environments is defined as any factor that hinders the accurate evaluation of the fitness of a given trial design. These factors can include the use of approximate fitness functions, the use of noisy data, knowledge uncertainty, sampling and human error. A GA that operates in a noisy environment is referred to as a "Noisy GA". Noisy GAs were used for the first time for image registration [Grefenstette and Fitzpatrick, 1985; Fitzpatrick and Grefenstette, 1988]. They can also be used in problems where the optimal design must be effective for a wide range of parameter values.

A type of noisy fitness function called a "sampling fitness function" is often used in noisy GAs [Miller and Goldberg, 1996]. This function uses sampling in order to evaluate the amount of noise from fitness evaluations in noisy environments. Sampling is performed by taking the mean of multiple noisy fitness evaluations for a given trial design in accordance with the Central Limit Theorem. However, this could result in an entire range of possible fitness functions being generated by simply changing the sample size of the sampling fitness function. Thus, different sample sizes produce fitness functions that are pareto-optimal in terms of speed and accuracy [Miller and Goldberg, 1996]. As time is an important constraint in solving real world problems, determining an effective sample size would considerably reduce the computational effort involved. This paper develops a methodology that will enable the identification of an optimal sampling strategy for a fixed computational time using theoretical relationships from the genetic algorithm literature. This method is applied to a groundwater remediation design problem.

## BASICS OF A NOISY GENETIC ALGORITHM

A GA searches a decision space using methods that are analogous to Darwinian "natural selection" processes. The decision variables associated with the problem are encoded as binary 0-1 digits. The fitness of each trial design is evaluated using a user-defined objective function for the problem of interest. After each individual is assigned a fitness value, the GA solves the optimization problem by using the basic Darwinian operators – selection, crossover and mutation [see Goldberg (1989) for more background on the simple GA].

When a noisy GA is used, a "noisy" fitness value $f_i$*'* for an individual *i* is returned that is the sum of the real fitness of the individual, $f_i$ and a random noisy component. As shown by *Miller (1997),* the noisy fitness value of an individual *i* can be described by the equation

$$f_i' = f_i + noise \qquad (1)$$

As the noise component is randomly drawn in each evaluation, subsequent evaluations of the same individual return different noisy fitness values [Miller and Goldberg, 1996].

As shown by *Miller and Goldberg (1996),* the evaluation of an individual using a sampling fitness function with a sample size of *n* can be expressed as

$$f_{i,n}{}^{*} = \frac{1}{n}\sum_{j=1}^{n} f_{i,j}^{*} \qquad (2)$$

where $f_{i,j}{}^{*}$ is the *j*th noisy fitness evaluation of individual *i*.

Unlike the Monte Carlo simulation modeling which requires extensive sampling, the noisy GA with the sampling fitness function performs best with few samples. Goldberg (1989) states that the optimal solutions in a GA are obtained by combining highly fit building blocks in the populations of strings. As the population contains many strings representing each building block, multiple samples of a particular building block's fitness will be found in the population, even if only one sample is drawn from the noisy fitness function for each string. As the population evolves, only those strings that have the highest fitness under all sampled conditions will be able to dominate and hence the noisy GA gives robust designs.

When determining the optimal sample size for the sampling fitness function, the tradeoffs between increasing computational time and decreasing the noise level are considered. Using a larger sample size ensures that the noise is reduced (as seen from equation 2) but results in additional computational time for the fitness function. The optimal sample size is where the performance penalty due to additional sampling is balanced by the faster convergence of the GA due to lower noise variance. The following sections describe how an optimal sampling strategy for the noisy GA can be determined for real world problems.

## POPULATION SIZING AND PARAMETER SETTING

The most important step in designing a competent noisy GA is fixing the population size correctly. This is especially true for cases where the GA must find the optimal solution in a fixed amount of time. When the population size is too large, redundant individuals are processed, which reduces the number of generations the GA can process in a fixed time and hence reduces the solution quality. On the other hand, a small population size can cause the GA to converge prematurely to a sub-optimal solution. *Harik et al 1997* developed a population sizing model called the "random walk" based on the gambler's ruin problem. *Miller and Goldberg (1996)* modified this model to account for the presence of noise in the system (Equation 3).

$$N \geq -2^{K-1} \ln(\alpha) \left( \frac{\sqrt{\pi(\sigma_F^2 + \sigma_N^2 / n)}}{d} \right) \tag{3}$$

where N is the population size, K is the building block order, d is the minimum signal difference between competing individuals, $\sigma_F^2$ is the variance of the true fitness function, $\sigma_N^2$ is the variance due to the noise and n is the sample size. The variance of the true fitness function $\sigma_F^2$ describes the variance that would be determined if there was no uncertainty in the system. The variance of the noise $\sigma_N^2$ describes the variance of the fitness of each design when sampling is done, i.e the design is exposed to a wide variety of conditions. The three-step method developed by *Reed et al (2000)* is used as the starting point to determine the population size that will result in optimal performance of the GA. The primary difference between the equation used here and the model used in *Reed et al (2000)* is the presence of the term $\sigma_N^2/n$ relating the effect of noise on population size. The evaluation of this additional term is necessary in order to size the population adequately.

As mentioned in *Miller (1997)*, the sum of the population fitness variance $\sigma_F^2$ and the noise variance $\sigma_N^2$ can be assumed to be equal to the initial noisy fitness variance of the population. *Reed et al, 2000* used a randomly generated trial population with 1000 members to determine the variance and showed that this resulted in a conservative estimate of the population size. Similarly, a trial population with 1000 individuals and a sample size of 1 is used to determine the variance of the noisy fitness function. As can be seen from Equation 3, using a larger sample size results in a smaller value of the variance and hence a lower value of the population size. Hence, using a sample size of unity in the trial population results in conservative estimates for the population size.

However, determining the noise variance $\sigma_N^2$ for a given fitness function is necessary in obtaining the optimal sample size. The noise variance for a given fitness function can be set to the average mean noise variance of the trial population. This is usually used when the noise component is dependent on the fitness of the individual. This can be obtained by selecting *x* individuals, using *y* samples to obtain the noise variance of each individual and then taking the mean of the *x* noise variances that were obtained [Miller, 1997]. For cases where the noise component is not dependent on the fitness of the individual, the noise variance can be obtained by determining the variance of *y* samples of a randomly selected individual in the trial population.

The value of the true fitness variance can then be obtained by subtracting the variance due to the noise from the variance of the noisy fitness function determined earlier.

Tournament selection is the selection procedure used in this research. As recommended by *Reed et al (2000)*, the probability of crossover for tournament selection is set at 0.5 and the probability of mutation ($P_m$) is set at

$$P_m = \frac{1}{N} \tag{4}$$

Once the population size is estimated, a range of sample sizes can be determined using theoretical relationships from the literature. However, determining the optimal sample size is difficult as each sample size produces a fitness function that is pareto-optimal in terms of speed and accuracy. *Miller (1997)* presented methods for identifying lower and upper bounds on the sample size. By bounding the range of sample sizes considered for the sampling fitness function in this manner, the computational effort involved can be significantly reduced.

## ESTIMATING THE LOWER BOUND OF THE SAMPLE SIZE

In order to determine the lower bound of the optimal sample size for a computationally constrained GA, the computational requirements of the GA need to be modeled over time. *Fitzpatrick and Genfenstette (1988)* developed a model for the total time T required by the GA as

$$T = (\alpha + \beta * n)GN \tag{5}$$

where $G$ is the total number of generations, $N$ is the population size and n is the sample size of the sampling fitness function. The variable $\alpha$ represents the fixed amount of GA overhead time per individual per generation, which includes the time required for selection, crossover and mutation but not for the fitness function evaluation. The variable $\beta$ represents the time required for a single fitness function evaluation. The costs of generating the initial population have been ignored as they are negligible when compared to the costs of running the GA.

The above equation can be used to determine the value of the ending generation as a function of T, $\alpha$, $\beta$, $n$ and $N$. From Equation 3, it can be seen that the population size $N$ is also a function of the sample size. Hence, the value of the ending generation reduces to a function of the sample size $n$ as all of the other values are assumed constant. An important assumption used by *Miller and Goldberg (1996)* to develop the lower bound for the sample size is that the GA runs for all sample sizes will have the same convergence rate. Hence, the optimal sample size is the sample size that maximizes the ending generation. *Miller and Goldberg (1996)* used this to show that the lower bound for the optimal sample size ($n_{lb}$) can be given by

$$n_{lb} = \sqrt{\frac{\alpha}{\beta}} \sqrt{\frac{\sigma_N^2}{\sigma_F^2}} \tag{6}$$

where $\sigma_F^2$ is the true fitness variance and $\sigma_N^2$ is the noise variance. The values for the variances were determined earlier and the other variables $\alpha$ and $\beta$ can be calculated through a trial run of the noisy GA. Using these values, the lower bound of the sample size can be determined. It in interesting to note at this point that the lower bound is domain independent.

## ESTIMATING THE UPPER BOUND OF THE SAMPLE SIZE

*Miller (1997)* suggested developing an approximate convergence model for the fitness function in order to be able to estimate an upper bound of the optimal sample size. As the selection pressure is the primary component that determines the convergence rate of the GA, selection pressure convergence models can be used to develop approximate convergence models. *Miller (1997)* describes an approximate convergence model that is representative of GA convergence in most domains as

$$\overline{f}(t) = \frac{e^{\frac{x}{\sigma_N}t+c}}{1+e^{\frac{x}{\sigma_N}t+c}} \tag{7}$$

where *x* is a function of the convergence rate, *c* determines the starting population fitness, *t* is the time and *f(t)* represents the fitness as a function of time.

This model can be calibrated by determining the initial value of the fitness and the final value of the fitness at the end of the GA's run in order to compute *x* and *c*. While calculating the upper bound of the sample size, the value of the fitness at a time *t* equal to the ending generation is determined. This results in the equation

$$p_{ub} = p(G) \tag{8}$$

where pub is the fitness at the upper bound of the sample size and G is the ending generation which is a function of the sample size.

As shown by *Miller (1997)*, the upper bound is then determined by maximizing the performance of the GA ( $\frac{dp_{ub}}{dn} = 0$) and solving for the value of the upper bound (*$n_{ub}$*).

This method has been used in a case study to determine an effective sampling strategy so that uncertainty can be incorporated into the problem. An optimal solution is to be found that is effective under all the values of the parameters developed in the problem.

## THE CASE STUDY

A case study was developed using data derived from the Borden site as given in *Smalley et al (2000)*. The aquifer configuration is shown in Figure 1 and was modeled using a coarse grid of 16 by 8 elements. This coarse grid was derived from a finer mesh that was used to generate conditional hydraulic conductivity realizations as shown in *Smalley et al (2000)*. Benzene with an initial peak concentration of 133 mg/L was assumed to be present on the site. Multiple parameter sets were defined with each set consisting of a single sample drawn randomly from the pool of generated hydraulic conductivity realizations and from each of the nine variable exposure model parameter distributions.

The remediation strategy was to use extraction wells from two possible sets of well locations to remediate the aquifer. Pump and treat was the treatment of choice in this case in order to minimize the computational effort involved in evaluating the fitness function. Monitoring was performed at the locations specified in Figure 1.

The goal of the optimization is to identify an effective pump and treat remediation strategy that will meet a target risk at the least cost. An existing simulation model RT3D (see *Clement et al, 1998* for details) was used to simulate the effect of a remediation strategy. The fitness of each member of the population is determined by a penalty based objective function that includes violations in meeting the target risk and head constraints present in the system.

The “noise” in the system resulted from the uncertainty in the values of the hydraulic conductivity at the site as well as the presence of nine variable exposure parameters.

## RESULTS

The case study was run with several sample sizes in order to establish the optimal sampling strategy in this case. The results from this research will be presented at the conference.

## REFERENCES

Clement T.P, Y. Sun, B.S Hooker and J.N Petersen, Modeling MultiSpecies Reactive Transport in Ground Water, *Ground Water Monitoring and Remediation*, 18(2), 79-92, 1998.

Fitzpatric J.M and Grefenstette J.J, Genetic algorithms in noisy environments, *Machine Learning*, 3, 101-120, 1988

Goldberg, David E., *Genetic Algorithms in Search, Optimizations and Machine Learning*, Addison –Wesley, New York, NY, 1989.

Grefenstette J.J and Fitzpatric J.M, Genetic search with approximate function evaluations, In Grefenstette, J.J (Ed.), *Proceedings of an International Conference on Genetic Algorithms and their Applications*, pp 112-120, Hillsdale, NJ, 1985.

Harik, G.R, E. Cantu-Paz, D.E Goldberg and B.L. Miller, The gambler’s ruin problem, genetic algorithms and the sizing of populations, In *Proceedings of the 1997 IEEE Conference on Evolutionary Computation*, pp 7-12, IEEE press, New York, NY, 1997.

Miller B.L and D.E. Goldberg, Optimal Sampling for Genetic Algorithms, In Dagli, C.H, Akay, M., Chan, C.L.P Fernandez, B.R., and Ghosh J. (Eds.), *Intelligent Engineering systems through artificial neural networks* (ANNIE ’96), Volume 6, pp 291-298, New York, ASME Press, 1996

Miller B. L, *Noise, Sampling and Efficient Genetic Algorithms*, IlliGAL Report No. 97001, May 1997.

Reed Patrick, Barbara Minsker and David Goldberg, Designing a Competent Simple Genetic Algorithm for Search and Optimization, *Water Resources Research*, 36(12), 3757-3761, 2000.

Ritzel, Brian J.,  Eeheart J. W. and S. Ranjithan, Using genetic algorithms to solve a multiple objective groundwater pollution containment problem, *Water Resources Research*, 30(5), 1589-1603, 1994.

Smalley, J. B., B. S. Minsker, and D. E. Goldberg, Risk-based in situ bioremediation design using a noisy genetic algorithm, *Water Resources Research*, 36(20), 3043-3052, 2000

Wang Q.J, The genetic algorithm and its application to calibrating conceptual rainfall-runoff models, *Water Resource Research*, 27(9), 2467-2471, 1991.