Adaptive Sampling for Noisy Problems

Erick Cantú-Paz

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
Livermore, CA 94551
cantupaz@llnl.gov

Abstract. The usual approach to deal with noise present in many real-world optimization problems is to take an arbitrary number of samples of the objective function and use the sample average as an estimate of the true objective value. The number of samples is typically chosen arbitrarily and remains constant for the entire optimization process. This paper studies an adaptive sampling technique that varies the number of samples based on the uncertainty of deciding between two individuals. Experiments demonstrate the effect of adaptive sampling on the final solution quality reached by a genetic algorithm and the computational cost required to find the solution. The results suggest that the adaptive technique can effectively eliminate the need to set the sample size a priori, but in many cases it requires high computational costs.

1 Introduction

Evolutionary algorithms (EAs) are considered relatively robust to the noise present in the evaluation of the objective function of many real-world optimization problems [1]. The usual approach to deal with noise is to take an arbitrary number of samples of the objective function and use the average as an estimate of the true objective value. If many samples are taken, the estimates will be very accurate, but it may be a waste of computing resources. On the other hand, if too few samples are taken, the algorithm may incorrectly select inferior solutions and this might lead to failure. This paper presents results with an adaptive sampling technique that adjusts the number of samples to the uncertainty in deciding between two specific individuals. This adaptive approach eliminates the need to selecting the number of samples a priori.

The objective of this paper is to study the effect of adaptive sampling on the final solution quality and execution time. This paper presents a systematic study of the impact of the sampling using a simple test problem (a 100-bit onemax function). The experiments demonstrate that the adaptive sampling can find better solutions than an arbitrary fixed sample and it can save time over an excessively large fixed sample.

An alternative way to deal with noise in the fitness evaluations is to increase the population size [2, 3]. The paper shows that increasing the population size seems a very effective and efficient way to deal with noise. However, in practice

it may be difficult to determine the correct population size because current population sizing models require knowledge of problem-specific parameters and the noise intensity [2, 3]. If the parameters are unknown or if the noise levels are not uniform, it will be difficult to use these models accurately.

The next section provides some background on previous work in solving noisy problems with EAs. Section 3 presents in detail the adaptive sampling method. Section 4 shows the results of experiments. Finally, section 5 concludes the paper and discusses future work.

2 Background

The robustness of evolutionary algorithms to noise in the evaluation of solutions has been recognized for a long time. Recent research suggests that the use of populations is the cause of the robustness of EAs in noisy environments [1]. Harik et al. [2] presented models to determine the size of the populations required to solve certain types of problems and considered the case where the fitness evaluations are noisy. Miller [4] extended Harik et al.'s model to account for sampling the objective function, and later presented models to optimize the sample size [3].

There is still controversy about the tradeoff between increasing the sample size or increasing the size of the populations. Sampling the objective function n times increases the computation time by a factor of n, but reduces the standard deviation of the estimate by a factor of only $\sqrt{n}$. Fitzpatrick and Grefenstette [5] argue in favor of increasing the population size rather than the sample size. Arnold and Beyer's [6] calculations for the sphere model agree that increasing the population is beneficial in evolution strategies with intermediate recombination. On the other hand, Beyer [7] argues that in a $(1, \lambda)$-ES, the sample size should be increased, rather than λ. Hammel and Bäck [8] verify Beyer's result and show that there is no benefit of increasing the parent population size.

The previous works assume that the sample size is fixed beforehand and remains constant during the execution of the EA. Aizawa and Wah [9] were probably the first to introduce a method that allocates different number of samples to different individuals. Their objective is to find an allocation that minimizes a pre-defined loss function. They minimize the expected estimation error as the loss function, which has the effect of drawing more samples from better individuals and spending less time in the inferior ones. Branke and Schmidt [10] also proposed an adaptive sampling method that takes additional samples of both individuals participating in a tournament until the normalized fitness difference between the two individuals falls below some threshold. The normalized fitness difference is obtained dividing the difference of the observed fitnesses by the standard deviation of the difference: $(\bar{f}_x - \bar{f}_y)/\sigma_d$.

The approach presented in the present paper differs from Aizawa and Wah's in that our objective is to take the smallest number of samples necessary to make a decision between competing individuals during the selection process. Our approach is very similar to Branke and Schmidt's, but differs in that we take

samples one at a time from the individual with the highest observed variance, and we use standard statistical tests to select the winner of the tournament with certain probability. In addition, Branke did not examine the impact of the technique on the final solution quality, only on the probability of selecting the correct individual.

Somewhat similar to our approach, Teller and Andre [11] proposed a method that allocates varying numbers of fitness cases to evaluate individuals in genetic programming. With their algorithm, individuals are initially evaluated on a small number of fitness cases, and are further evaluated only if there is some chance that the outcome of the tournaments they participate in can change. There is no point on refining evaluations of individuals that are so much better (worse) than their competitors that they are not likely to lose (win) their tournaments. A similar algorithm was developed independently by Giacobini et al. [12].

3 Adaptive Sampling

We consider pairwise tournament selection, where the best of two randomly chosen individuals is selected to continue in the algorithm. Without loss of generality we consider maximization problems. The noisy fitness F' of an individual can be described as

$$F' = F + N, \tag{1}$$

where F is the true fitness and N is the added noise. In this paper, we use normally distributed noise: $N \sim N(0, \sigma_N^2)$, but the same approach can be used with other noise distributions.

Assume that we want to compare two individuals x and y. In a noisy environment, their fitnesses are the random variables $F_x \sim N(\mu_x, \sigma_x^2)$ and $F_y \sim N(\mu_y, \sigma_y^2)$. If $\mu_x > \mu_y$, we would like to select individual x. However, the true means are unknown and we estimate them using the averages $\bar{f}_x = \frac{1}{n_x} \sum_i^{n_x} f_i$. and $\bar{f}_y$ of multiple samples of the fitness function. The true variances are also unknown, so we approximate them with the observed variances s_x^2 and s_y^2. When dealing with noise, it is common to use an arbitrary number of samples and choose the individual corresponding to the highest mean. However, this simple approach may lead to choosing the wrong individual, because it does not take into consideration the uncertainty in the estimations. If too few samples are taken, the estimates will be inaccurate and may lead to failures. If too many samples are taken, computational resources will be wasted.

Let $d = \bar{f}_x - \bar{f}_y$ be the difference between the observed means of individuals x and y. By the central limit theorem, as the number of samples increases the distributions of $\bar{f}_x$ and $\bar{f}_y$ approach normal distributions, regardless of the noise distribution. The distribution of d also approaches a normal $d \sim N(\bar{f}_x - \bar{f}_y, s_x^2 + s_y^2)$. The probability P_c of choosing the individual with the highest quality is

$$P_c = \Phi\left(\frac{\bar{f}_x - \bar{f}_y}{\sqrt{s_x^2 + s_y^2}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Note that deciding correctly between the two individuals becomes more difficult as d becomes smaller and the observed variances become larger.

The proposed adaptive sampling method is simply to estimate the mean fitnesses and variances using a small number of samples initially and then sample the fitness of the individual with the highest variance until a statistical test can decide between the individuals with some certainty or a limit in the number of samples is reached. There are many possible variations on this idea depending on the initial number of samples, and the way to determine the certainty of the decision. Branke and Schmidt [10] used 10 initial samples of each individual and sample both individuals until the normalized fitness difference falls below some user-specified threshold. We test the method taking only two initial samples from each individual. Instead of using an arbitrary threshold, we use a statistical test to determine when to stop sampling. Branke and Schmidt studied the effect of sampling in the probability of selecting between two individuals, but did not present results of the impact of the adaptive sampling on the final solutions or the cost associated with finding those solutions. We extend their study in those directions.

For the experiments in this paper, we take only two initial samples from each individual. Then the means and variances of the fitness of each individual are estimated. As additional samples are taken, the means will approach a normal distribution, but since we have a small number of initial samples, we decide between the two individuals using a one-sided t test. We conservatively use the minimum of the number of samples of x and y as the degrees of freedom in the test. If the p value of the test is greater or equal to (an arbitrarily chosen) 0.9, then we consider that the test discriminates between the individuals with sufficient certainty. If the p value is below our threshold, then we take an additional sample from the individual with the highest variance and repeat the test.[1] Resampling continues until the p value meets the threshold.

Figure 1 shows the probability that the adaptive sampling procedure chooses the best individual as a function of the difference of their fitnesses. The experiments consisted of using the adaptive sampling to compare an individual with real fitness of 100 to an individual with fitness $100 - i$ for $i = 1, ..., 20$ adding unbiased normal noise with standard deviation of 5, 10, and 20. For each fitness difference we execute 10000 trials and report the fraction of trials where the individual with true higher fitness was selected along with 95% confidence intervals. If the fitnesses of the two individuals are equal, the decision is a random choice. As the fitness difference increases, the probability of deciding correctly quickly approaches 1.0. This suggests that it is easy to decide between individuals with large fitness differences and/or low variance. Therefore we should spend additional computational resources in dealing with individuals with small differences

[1] Strictly, re-testing using the same samples might lead to elevated type-I errors and sequential testing methods might be required. However, as the experiments demonstrate (see figure 1), the proposed procedure chooses the correct individual very consistently, so we opted for the simpler re-testing.

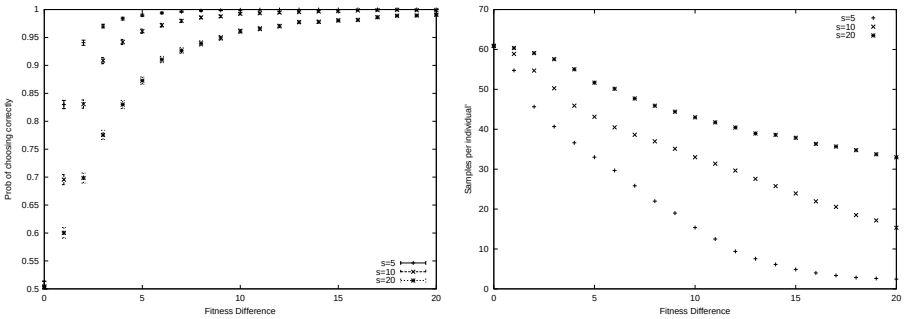


Fig. 1. The probability of correctly choosing the best individual using the adaptive sampling method (left) depends on the difference of the fitnesses of the two individuals. The right graph displays the number of samples required by the adaptive sampling.

and/or large variances. The figure also shows the number of samples necessary to decide correctly with the proposed technique.

The next section examines the effect of using this method on the final solution quality as well as the computational cost. We examine the effect of varying different parameters of interest, such as the noise levels, the population size, and the thresholds for the p values.

4 Experiments

The experiments use a simple generational GA with uniform crossover applied with probability 1.0. No mutation was used to try to limit the source of randomness to the exogenous noise added to the fitness function. The GA used pairwise tournament selection without replacement. The population sizes and noise levels are indicated in each experiment. The GA was terminated when all the members of the population were identical. The random number generator was a Mersenne Twister [13] initialized with 32 bit unsigned integers obtained from www.random.org. All results presented are over 100 repetitions of each parameter setting, and the graphs include 95% confidence intervals.

The results of the adaptive sampling method will be compared against a GA that always uses 10 samples to estimate the objective value and against a GA that ignores the noise and selects tournament winners based on a direct comparison of a single function evaluation.

4.1 Noise and Selection Intensity

A way to measure progress in solving a problem is to examine the average fitness of the population as a run progresses. The increase in average fitness depends on the intensity of the selection method used. Miller suggested that noise reduces the selection intensity [14]. Selecting incorrectly the individuals that have

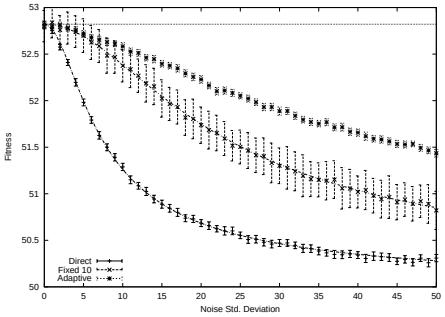


Fig. 2. Effect of noise on the average (real) fitness of individuals selected by pairwise tournament selection. The adaptive method remains closer to the expected fitness than the fixed-size sampling.

lower fitness values has the effect of reducing the average fitness of the selected individuals. The selection intensity is defined as

$$I = \frac{\mu_{\text{sel}} - \mu_F}{\sigma_F},$$

where μ_{sel} is the mean fitness of the selected individuals and μ_F and σ_F are the mean fitness and standard deviation of the original population. Each selection method has a different intensity, and for pairwise tournament selection $I = 0.5642$. Miller [3] established that the expected fitness of the selected individuals in a noisy environment is:

$$\mu_{\text{sel}} = \mu_F + I \frac{\sigma_F^2}{\sqrt{\sigma_F^2 + \sigma_N^2}}. \tag{3}$$

This result can be easy extended to account for the reduction in the uncertainty that comes from taking a number n of multiple samples:

$$\mu_{\text{sel}} = \mu_F + I \frac{\sigma_F^2}{\sqrt{\sigma_F^2 + \sigma_N^2/n}}. \tag{4}$$

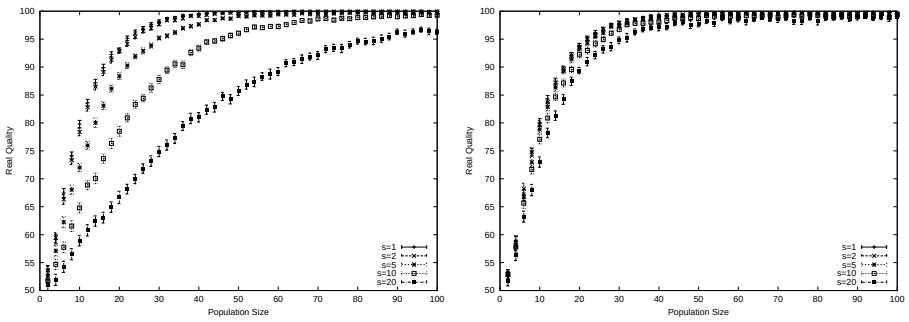
Figure 2 shows the effect of noise on the mean fitness of the individuals selected by pairwise tournament selection. To obtain these results a population of 1000 individuals was initialized randomly and we measured the average fitness of the individuals selected using tournament selection. The continuous lines in the figure were obtained with equation 4. The graph shows that ignoring the noise (using one sample) can reduce the selection intensity quite strongly. With 10 fixed samples, the reduction is much less severe as expected. The best results are obtained with the adaptive sampling.

It is not clear, however, what will be the effect of the adaptive sampling on the overall performance of the algorithm. The next two subsections examine this.

4.2 Noise and Solution Quality

Figure 3 shows the means of the real fitnesses at the end of the runs of a GA that uses direct comparisons and a GA with the adaptive sampling method. The experiments considered different noise levels and the population size was varied between 2 and 100 individuals. As expected, the figure shows that the final solution quality improves with larger populations, but with direct comparisons the quality degrades as the noise level increases. The adaptive sampling method shows much smaller quality degradations that are significant at the 0.95 significance level only with high noise strengths of $\sigma = 10$ and 20.

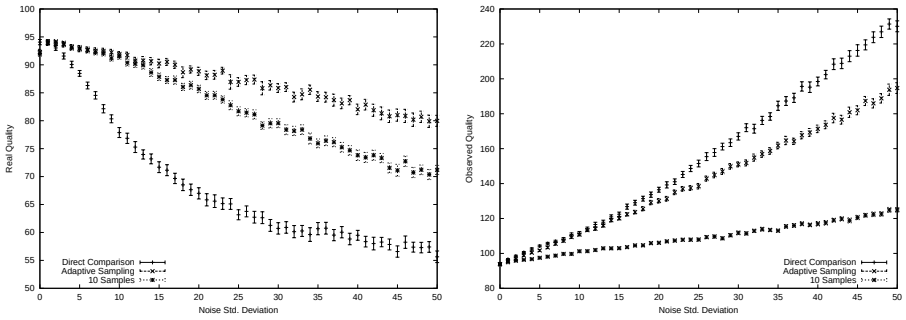
The effect of noise in the final quality of solutions can be observed more directly in figures 4 and 5. In these figures, the population size is kept constant at $N = 20$ and $N = 40$ individuals while the noise level is varied. The left panels show the *observed* fitness values at the end of the runs and the left panels show the *real* fitness. As expected, ignoring the noise and making direct comparisons based on one sample has the worst results: The algorithm is misled with very large apparent fitness values that belong to individuals with very low real fitnesses. The adaptive method always finds the solutions with the best real fitnesses and appears much more robust to increased noise levels than the other two methods (i.e., the real fitness decreases much less with increasing noise).



(a) Direct (b) Adaptive

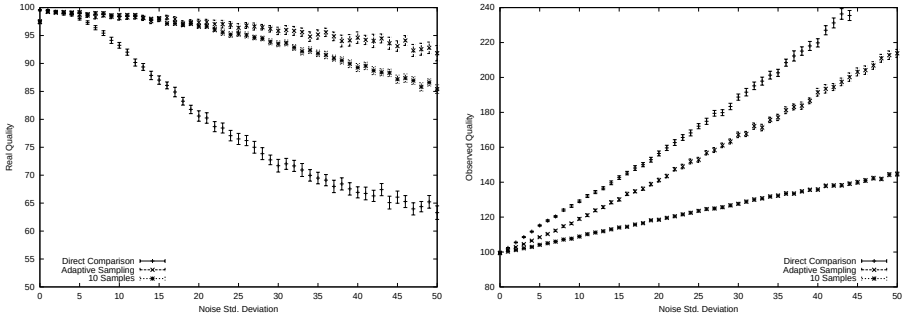
Fig. 3. Final real fitnesses for different noise levels varying the population size. The quality degradation is much smaller in the adaptive algorithm. Error bars denote 95% confidence intervals.

In the adaptive method, the user does not have to specify the number of samples per individual, but has to specify the p value of the test. The results in figure 6 show that the final solution quality is not affected greatly for a range of commonly used values of p, and the assignment $p = 0.9$ used in the experiments seems an appropriate choice.



(a) Real (b) Observed

Fig. 4. Real and observed fitnesses for a 100-bit onemax problem for different noise levels and a population of 20 individuals.



(a) Real (b) Observed

Fig. 5. Real and observed fitnesses for a 100-bit onemax problem for different noise levels and a population of 40 individuals.

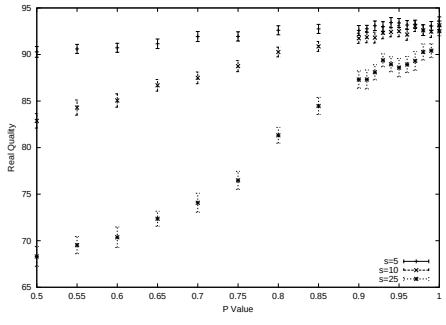


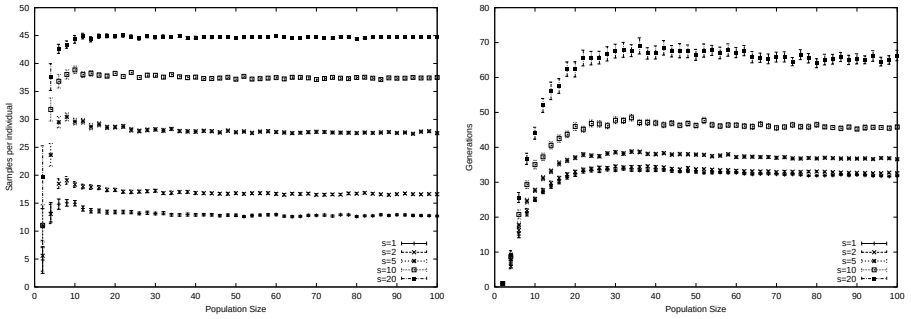
Fig. 6. Mean real quality for different noise levels varying the parameter p.

4.3 Computational Cost

So far we have seen that adaptive sampling can eliminate the need to fix the number of samples without sacrificing solution quality. However, eliminating the guess of the right sample size comes at a significant computational cost.

Figure 7 shows the means of the number of samples per individual and the number of generations until convergence of the experiments corresponding to figure 3. The number of samples appear to be dependent of the noise level and independent of the population size, except for very small populations. Similar observations can be made for the number of generations.

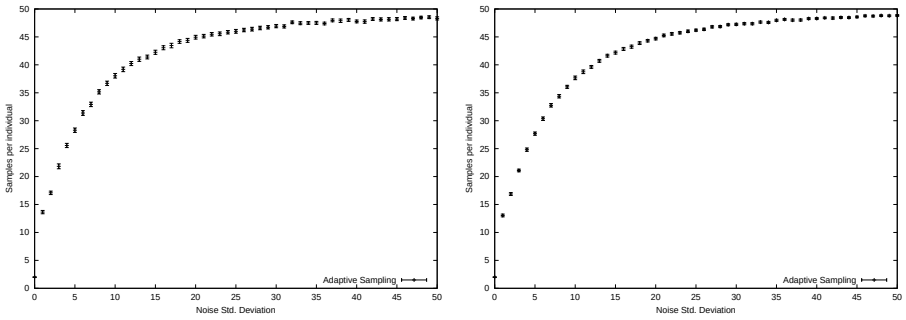
While the results of the previous subsection present a favorable outcome of the sampling methods, a fairer comparison of these algorithms should take into account the total computational cost. The total number of function evaluations taken by each experiment is the product of the population size, the number of generations, and the number of samples. Figures 8 and 9 present the number of samples and generations corresponding to the experiments in figures 4 and 5. Figure 10 compares the total cost incurred by each algorithm to reach solutions of a particular quality. The results show that the sampling methods need an order of magnitude more computations to reach the same solutions than simply ignoring the noise and using larger populations.



(a) Samples (b) Generations

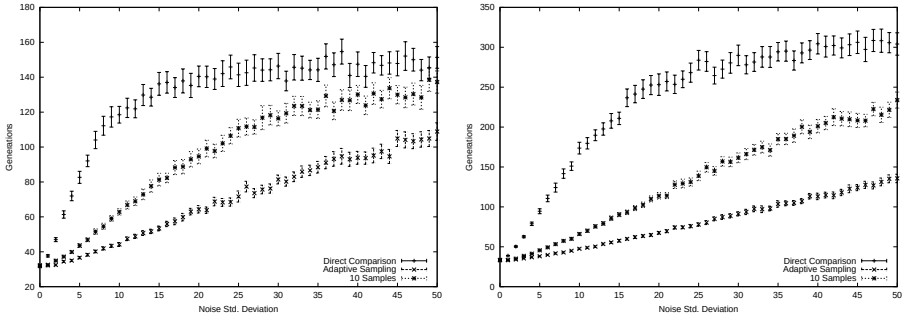
Fig. 7. Mean number of samples and generations for different noise levels and varying the population size. The results shown are of the adaptive sampling algorithm.

As the run progresses we can expect the individuals to become more alike each other, reducing their fitness differences and making the decisions more difficult. Albert and Goldberg studying a different but related problem conclude that the sample size should increase over the run [15]. The results in figure 11 confirm this recommendation. We tracked the average number of samples used per generation in a problem with $N = 20$. In all cases, there is a clear upward tendency in the number of samples over time.



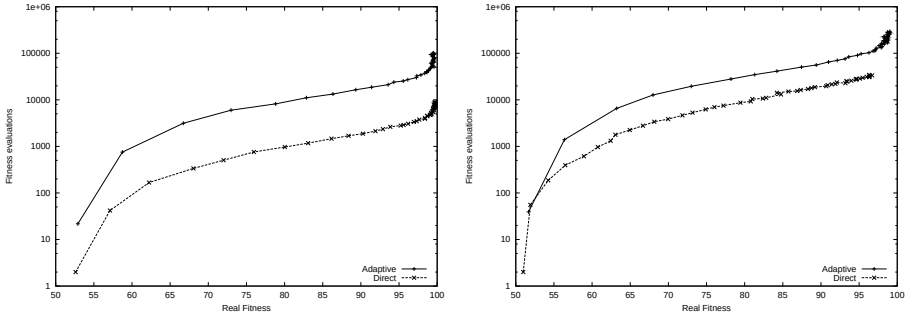
(a) $n = 20$ (b) $n = 40$

Fig. 8. Mean number of samples per individual varying the noise levels.



(a) $n = 20$ (b) $n = 40$

Fig. 9. Mean number of generations until termination varying the noise levels.



(a) $\sigma = 5$ (b) $\sigma = 20$

Fig. 10. Number of fitness evaluations required to reach different solution qualities.

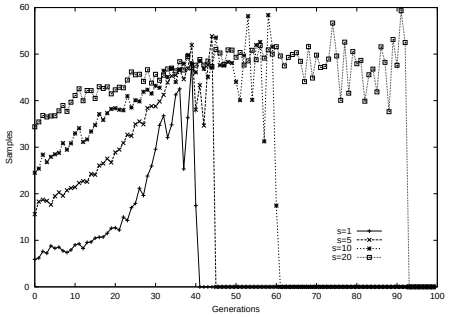


Fig. 11. Number of samples per generation under different noise levels. $N = 20$.

5 Conclusions

Adapting the number of samples to the uncertainty of the decision between two specific individuals eliminates the need to set the sample size a priori. We demonstrated that the adaptive method used in this paper can adapt to a large range of noise levels. For a fixed population size, the experiments showed that the quality reached with adaptive sampling is much higher than ignoring the noise or taking too few samples. The method was also demonstrated to be fairly robust to its only parameter, the p value of the test. For commonly used p values (0.85–1.0), the final quality does not differ much.

The main drawback of the method is that the convenience of adapting the sample size requires a substantial amount of computations. The experiments suggest that increasing the population size slightly and using direct comparisons (no resampling) might be the best strategy. However, the adaptive method is useful in practice, where there might be no knowledge of the noise level or the domain-dependent parameters necessary to use existing population sizing models. Also, in some applications the noise varies over time or the noise may not be uniform in the entire search space, and adapting the sample size will be especially beneficial in those situations.

Future work should include a more detailed investigation of the adaptive sampling with other test functions and adding highly biased noise. Although the assumption of normality is warranted for sufficiently large numbers of samples, it may be possible to mislead the algorithm if the observed fitness difference is small because very few samples will be taken. It is not clear what will be the effect of these mistakes on the overall performance of the algorithm.

Acknowledgments

UCRL-CONF-203216. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

References

1. Arnold, D.V., Beyer, H.G.: A comparison of evolution strategies with other direct search methods on the presence of noise. Computational Optimization and Applications **24** (2003) 135–159
2. Harik, G., Cantuú-Paz, E., Goldberg, D.E., Miller, B.L.: The gambler's ruin problem, genetic algorithms, and the sizing of populations. Evolutionary Computation **7** (1999) 231–253
3. Miller, B.L.: Noise, sampling, and efficient genetic algorithms. doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana (1997) Also IlliGAL Report No. 97001.
4. Miller, B.L., Goldberg, D.E.: Optimal sampling for genetic algorithms. Proceedings of the Artificial Neural Networks in Engineering (ANNIE '96) conference **6** (1996) 291–297
5. Fitzpatrick, J.M., Grefenstette, J.J.: Genetic algorithms in noisy environments. Machine Learning **3** (1988) 101–120
6. Arnold, D.V., Beyer, H.G.: Local performance of the $(\mu/\mu_i, \lambda)$-ES in a noisy environment. In Martin, W., Spears, W., eds.: Foundations of Genetic Algorithms, Morgan Kaufmann (2000) 127–142
7. Beyer, H.G.: Toward a theory of evolution strategies: Some asymptotical results from the $(1,^+ \lambda)$-Theory. Evolutionary computation **1** (1993) 165–188
8. Hammel, U., Bäck, T.: Evolution strategies on noisy functions: How to improve convergence properties. In Davidor, Y., Schwefel, H.P., Männer, R., eds.: Parallel Problem Solving fron Nature, PPSN III, Berlin, Springer-Verlag (1994) 159–168
9. Aizawa, A.N., Wah, B.W.: Scheduling of genetic algorithms in a noisy environment. Evolutionary Computation **2** (1994) 97–122
10. Branke, J., Schmidt, C.: Selection in the presence of noise. In Cantú-Paz, E., Foster, J.A., Deb, K., Davis, D., Roy, R., O'Reilly, U.M., Beyer, H.G., Standish, R., Kendall, G., Wilson, S., Harman, M., Wegener, J., Dasgupta, D., Potter, M.A., Schultz, A.C., Dowsland, K., Jonoska, N., Miller, J., eds.: Genetic and Evolutionary Computation – GECCO-2003, Berlin, Springer-Verlag (2003) 766–777
11. Teller, A., Andre, D.: Automatically choosing the number of fitness cases: The rational allocation of trials. In Koza, J.R., Kalyanmoy, D., Dorigo, M., Fogel, D.B., Garzon, M., Iba, H., Riolo, R.L., eds.: Genetic Programming 97, San Francisco, CA, Morgan Kaufmann Publishers (1997) 321–328
12. Giacobini, M., Tomassini, M., Vanneschi, L.: Limiting the number of fitness cases in genetic programming using statistics. In et al., J.J.M., ed.: Parallel Problem Solving from Nature (PPSN VII), Berlin, Springer Verlag (2002)
13. Matsumoto, M., Nishimura, T.: Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. ACM Transactions on Modeling and Computer Simulation **8** (1998) 3–30
14. Miller, B.L., Goldberg, D.E.: Genetic algorithms, selection schemes, and the varying effects of noise. Evolutionary Computation **4** (1996) 113–131
15. Albert, L.A., Goldberg, D.E.: Efficient discretization scheduling in multiple dimensions. In Langdon, W.B., Cantú-Paz, E., Mathias, K., Roy, R., Davis, D., Poli, R., Balakrishnan, K., Honavar, V., Rudolph, G., Wegener, J., Bull, L., Potter, M.A., Schultz, A.C., Miller, J.F., Burke, E., Jonoska, N., eds.: GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, New York, Morgan Kaufmann Publishers (2002) 271–278