LOCAL SEARCH ISSUES FOR THE APLICATION OF A SELF-ADAPTIVE HYBRID GENETIC ALGORITHM IN GROUNDWATER REMEDIATION DESIGN

Felipe P. Espinoza*, Barbara S. Minsker** and David E. Goldberg***

* Department of Civil and Environmental Engineering, University of Illinois, 205 N. Matthews Ave., Urbana, IL 61801, email: fespinoz@uiuc.edu
** Department of Civil and Environmental Engineering, University of Illinois, e-mail:minsker@uiuc.edu
*** Department of General Engineering, University of Illinois, e-mail:deg@uiuc.edu

Abstract

Water resources management problems can be computationally intensive and improved methods are needed to allow solution of more complex applications. In this paper, we study a numerical algorithm designed to efficiently solve water resources management applications such as groundwater management problems. The algorithm is a combination of a simple genetic algorithm and a local search method and is called a self-adaptive hybrid genetic algorithm (SAHGA). The paper presents new ways to improve performance of this algorithm together with an analysis of different alternative local search algorithms. The paper also includes an analysis of the reduction in population size that is possible when using SAHGA relative to a simple genetic algorithm (SGA). The results show that the improved algorithm is more reliable and effective in solving the proposed problem, with average savings of 68% with respect to the SGA.

Introduction

Groundwater remediation design is a complex non-linear problem with high demand on computer resources for a proper solution. Previously, we proposed a solution technique based on simple genetic algorithms (SGA) coupled with local search to help overcome this computational burden [Espinoza et al., 2001]. This technique is called the self-adaptive hybrid genetic algorithm (SAHGA). This algorithm has the capability of selecting local search only when new information can be generated. This selection is based on the proper use of past information about performance. In this new step of the research, we study how to improve the original algorithm in order to maximize the local search effect over the different components of the problem: local search applied to pumping rates only, well locations only, or combinations of both parameters. The study also includes an analysis of the effects of different local search algorithms.

The paper is divided into 5 sections. The first section gives a brief description of the test problems. The second section presents the proposed modifications to the algorithm. The next section includes the study of different local search algorithms.

The fourth section presents the analysis of performance for different design scenarios. Finally, the last section of the paper gives conclusions and recommendations.

Test Problem

The performance of the algorithm was assessed in the design of a groundwater remediation system for a hypothetical aquifer similar to the case studied by Smalley et al. (2000) and Gopalakrishnan et al. (2003). The aquifer is heterogeneous, isotropic and confined. It is assumed to be contaminated with BTEX with a peak initial concentration of 200 mg/L. The problem considers minimization of the pumping cost in order to meet a specific human health risk target. The treatment technology is assumed to be pump-and-treat. The pumping system is designed with up to three extraction wells with a maximum capacity of 250 m³/day each. The location of these 3 wells can be selected among 58 predetermined locations. Among these 58 wells, 15 are preexisting and are used as observation wells. The treatment technology for the extracted water is assumed to be air stripping.

For this problem, the designs are represented by 4 decision variables: pumping well location, pumping rate, and well installation. The cost of each design is evaluated by the following objective function (Smalley et al. 2000):

$$\text{Min } C_{TOT} = C_{REM} + C_{MON} + C_{SYST} \tag{1}$$

where the total cost C_{TOT} consists of three components – $C_{REM,}$ which is the capital and operating costs for the wells; $C_{MON,}$ which is the cost of on-site monitoring; and $C_{SYST,}$ which includes additional capital and operating costs for the ex-situ treatment system. Extensive details on the remediation and monitoring cost are presented by Smalley et al. (2000). The final component (C_{SYST}) was evaluated from data obtained from RACER (1999), a parametric cost modeling system, as detailed by Gopalakrishnan et al. (2003).

The optimal design must satisfy human health risk limits, pumping rate limits, and limits on hydraulic drawdowns (see Smalley et al., 2000, for details).

Finally, the fitness of the design is evaluated by combining the cost (from Eq. (1)) with penalties for violations of the risk and head constraints. No penalties are necessary for pumping rate limits because the pumping rate is one of the decision variables and it is limited directly in the GA through its representation as a binary number. The final fitness equation (with linear penalty) is presented in Eq. (2):

$$\text{Fitness} = C_{TOT} + \omega_1 \cdot \text{Risk\_violation} + \omega_2 \cdot \text{Head\_violation} \tag{2}$$

where ω_1 and ω_2 are the penalty weights for the risk and head constraints respectively. For this case study, the values of ω_1 and ω_2 have been set to 1000. Note that for this application, fitness is minimized so that low fitness designs are optimal.

Self-Adaptive Hybrid Genetic Algorithm (SAHGA)

The SAHGA algorithm works with the following operators: local search frequency, probability of local search and number of local search iterations. The first operator, local search frequency, is the switch between global and local search. The second operator, probability of local search, is the probability that local search will be performed on each member of the SGA population in each generation where local search is invoked. Finally, the third operator, number of local search operations, represents the maximum number of local search iterations (LS) before local search is halted. SAHGA adapts in response to recent performance of the algorithm as it converges to the solution. In other words, the operators are used only when they can provide new information to the search. The adaptation process for the three parameters is different. The global-local switch is based on the evaluation of the ratio of the coefficient of variation of the fitness between two consecutive generations. Local search is performed when this ratio is greater than a specified local search threshold. The second parameter to adapt is the probability of local search. This probability is adapted by decreasing from the initial value at the beginning of every local search step. Finally, the number of local search steps is controlled by comparing the improvement attained in the local search step with the improvement because of global search. For more details on SAHGA, see Espinoza et al. (2001).

Local Search Algorithms. The local search operator attempts to find the best solution starting at a previously selected point, in this case a solution in the SGA population. For this paper, 5 local search algortithms were evaluated. These algorithms are:

Random Walk with Uniform Distribution (LS1): In general, the random walk is simply the movement from one point of the decision space to a new point randomly selected using a uniform distribution from a neighborhood around the starting point [Spitzer, 1964]. One iteration of this algorithm requires one fitness function evaluation.

Random Walk with Normal Distribution (LS2): This algorithm is similar to the uniform distribution discussed previously, but the change of location is evaluated with a normal distribution instead of a uniform distribution. For this reason, the points located near the starting point are more likely to be selected than those located closer to the boundary of the search area. Again, one iteration requires one function evaluation.

(1+1)-Evolutionary Strategy (LS3): This algorithm, proposed by Rechenberg (1973) and Schwefel (1975), randomly selects a new location using a normal distribution with variable standard deviation. The standard deviation changes following the so-called $1/5$ success rule based on the evaluation of success of the search. This algorithm also requires one fitness function evaluation per local search iteration.

Random Derivative (LS4): This algorithm randomly selects a search direction, and using this direction, the location of a new point is randomly selected. This algorithm needs 2 fitness function evaluation for every local search iteration (one for finding the coordinates of the new point using derivatives and one for evaluating its fitness).

Steepest Descent (LS5): The algorithm evaluates the new point following the direction of the gradient of the function at the starting point. This algorithm performs one function evaluation for every one of the decision variables of the particular test problem (to numerically evaluate gradient) and one function evaluation to evaluate the fitness of the new individual.

Population Size. The population size evaluation for SAHGA is based on the methodology proposed by Reed at al. (2000), modified to include the local search effect. Local search affects the search by reducing the standard deviation of the fitness population, since several members may have the same nearby local optimum. This reduction can be included in the population sizing relationship from Harik et al. (1997) as follows:

$$N \geq -2^{K-1} \ln(\alpha) \left(\beta \frac{\sigma_f}{d} \right) \tag{3}$$

In Eq. (3), K represents the building block (BB) order; α is the reliability (i.e., the probability that the GA finds the optimal solution); σ_f is the standard deviation of the fitness function; and d is the signal difference between the best and second best solution. The parameters σ_f and d are estimated using a large, random initial population. The building block order, K, is unknown but can be assumed to vary between 1 and 5 [Reed at al., 2000]. In their design methodology, K starts at 1 and is increased until the difference between optimal solutions from 2 consecutive runs is not significant and the solution has taken over at least 80% of the population.

To evaluate population size for the hybrid algorithm, the value of β must be estimated. The parameter β is the standard deviation reduction and it is evaluated as the average of the reduction for the first 5 iterations, as presented in Figure 1, because of most of the search is performed with no more than 5 local search iterations. With this information, the value of β was evaluated as 0.83, 0.80, 0.80, 073 and 0.66 for the 5 local search algorithms.

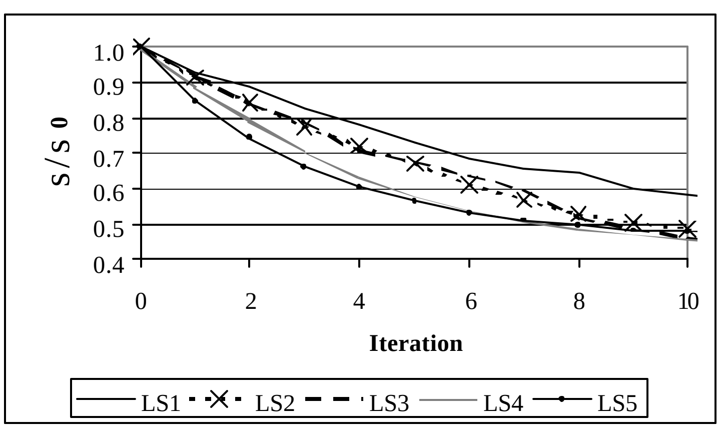

Figure 1. Standard Deviation Reduction

Finally, in order to evaluate the population it is necessary to estimate d, the signal difference between the best and the second best members of the population. In [Harik et al., 1997], d is defined as the difference in fitness of the best and second best building block in the population. For real applications, the building blocks are unknown. In this application, d was estimated from the differences in fitness of the best members of the population. In this case the analysis was performed by means of a probabilistic approach based on frequency analysis theory [Hogg and Craig, 1978], in which the histogram of the fitness function is evaluated. This histogram represents different "classes" of fitness that can be statistically identified. Using the histogram, "d" is evaluated as the size of the first class of the histogram. With all this information, the population for the SGA and the 5 local search algorithms is evaluated for different values of the building block order. Table 1 presents these results.

Table 1. Population Size

		Local Search Algorithm				
K	**SGA**	**LS1**	**LS2**	**LS3**	**LS4**	**LS5**
1	30	25	24	24	22	20
2	60	50	48	48	44	40
3	120	100	96	96	88	80
4	240	200	192	192	176	160
5	480	400	384	384	352	320

In order to reduce the computing time, we use an elitist serial solution process called "injection" to evaluate the optimal population size. In this approach, first we solve the problem for the base population (K=1). The next step is usually to solve the problem for the next population (K=2) starting from a random population [Reed et al., 2000]. In the new approach, this step is modified by randomly replacing one of the individuals in the new population with the best individual from the previous population. Using this approach with $\mu+\lambda$ selection, we guarantee that the search is always improving the solution, or in the worst case scenario the solution does not change. This process is an adaptation of the multi objective approach presented by Reed [2002].

Algorithm Enhancements for Improved Performance

This section of the paper develops improvements to the original SAHGA to improve local search performance without increasing the number of function evaluations. For this section, we use only uniform random walk local search with the population presented in Table 1 and the injection approach before mentioned. Three enhancements are investigated. The first is related to the definition of the stopping criterion for the number of local search iterations. The second modification is related to the selection of the individuals undergoing local search. The final modification is related to stopping the application of local search altogether.

Stopping Criterion. The first enhancement is related to the evaluation of the number of local search iterations. In SAHGA, a local search iteration continues when:

$$\frac{\Delta \text{Global}}{\text{pop}} < \frac{\Delta \text{Local}}{\text{fev}} \tag{4}$$

where ΔGlobal is the change in global search, ΔLocal is the change in local search, pop is the population size, and fev is the number of local search evaluations in the local search step. In Eq. (4), the value of ΔGlobal is evaluated as the difference between the best fitness between generations. Because this expression only takes into consideration the behavior of the best individual, it does not represent the population as a whole. For this reason, this paper will compare performing using the average fitness of the whole population with the original approach

Individual Selection. The second enhancement is related to the selection of individuals in the population for local search. In most hybrid search approaches, a certain percentage of the population is randomly selected for local search. Under the proposed enhancement, individuals are chosen using clustering that is tailored to the stage of the search. The GA search process for a constrained problem can be divided into three stages. In the first stage, all individuals in the population (or trial solutions) are usually infeasible, which means that all or some of the constraints defining the problem are violated. In the second stage, some of the individuals are feasible, and in the third stage most of the individuals are feasible. With this knowledge, the search can focus only on those aspects of the problem where the search can be most effective. The following sections present the proposed search process for each of these stages. At the end of the section, other modifications to the algorithm are presented that can be applied to all three stages of the search.

Stage 1: No Feasible Solutions: The first modification in this stage is related to the selection of the individuals undergoing local search. The current mechanism involves random selection of N individuals from the full population. In the new approach, the selection will be performed by means of Latin-Hypercube sampling [Iman et al., 1980] from clusters formed using three different criteria. The first criterion will be fitness, where the individuals will simply be ordered from the lowest to the highest fitness. Then the population will be divided into N groups with equal number of individuals (or as close to equal as possible). The second criteria will be similar to the first, but using total pumping rate instead of fitness. The third criterion will involve clustering the population into N clusters with similar well locations and pumping rates. Once the clusters are formed, one individual fom each group will be selected to undergo local search. In this way, the sampling assures that the individuals undergoing local search are probabilistically different because they are located in different parts of the search space, which allows a more thorough exploration of the space.

Stage 2: A Combination of Feasible and Infeasible Solutions: In the second stage of the generational process, a few individuals are feasible. In this case, the total population is divided between M_F feasible individuals and M_I infeasible individuals.

Using this information, the number of infeasible and feasible individuals undergoing local search is given by:

$$N_I = \left\langle \frac{M_I}{M_I + M_F} N \right\rangle, \qquad N_F = N - N_I \tag{5}$$

$$\text{where} : \langle x \rangle \text{ represents the nearest integer to } x$$

Eq. (5) allows for a distribution of sampling among feasible and infeasible solutions that is proportional to the total number of feasible and infeasible solutions in the population.

Stage 3: No Infeasible Solutions: This stage is completely analogous to the first one, and the selection of the individuals is performed in the same way.

End of Local Search. After careful examination of the results, it is clear that when the difference between the average value and the best value is small (usually less than 1%), local search does not help to solve the problem. For this reason, the third modification to SAHGA is to halt local search when this difference is small.

Evaluation of Algorithm Modifications. To evaluate the improvements proposed above, we performed 10 runs of each enhancement alone and then all three enhancements working together, with a different random seed for each set of runs. The results of the analysis are presented in Table 2. In this table, savings are evaluated by comparing the number of function evaluations necessary to find an optimal solution "close enough" (less than 0.2%) to the SGA solution. The table also includes the average speed up attained with the algorithm.

First, from the point of view of average savings, the best approach is to cluster by fitness (called "F order" in Table 2) together with the enhancement in the stopping criterion and the local search criterion. In fact, on average the new algorithm is more than 3 times faster than the SGA. The maximum savings are similar for all 8 SAHGA algorithms (the original and the enhancements). The minimum amount of savings show that most of the SAHGA algorithms were sometimes slower than the SGA. The original SAHGA algorithm was slower than the SGA for 50% of the runs, but for the enhancements this number is much lower. Only the fitness clustering approach combined with the enhancements in stopping criterion and local search criterion consistently outperformed the SGA for every realization, again showing the superiority of this approach.

A second analysis was performed to compare the savings attained for the original SAHGA algorithm and the fitness clustering approach for each run. The results are presented in Figure 2. From the results presented in Figure 1, it is clear that the improved algorithm is slower than the SAHGA for two of the runs, but unlike SAHGA is always faster than the SGA. This fact shows that the enhanced algorithm is more reliably faster than the original approach.

Table 2. Evaluation of Algorithm Modifications

Algorithm Enhancement	Savings relative to SGA			Average Speed Up
SAHGA Alone	-293.94	-23.40	85.39	-
F order	-35.75	61.73	87.47	2.61
Q order	-36.42	37.31	87.47	1.60
C order	-23.42	62.39	87.47	2.66
D**Global**	-78.13	43.65	88.09	1.77
End of Local Search	-534.56	-32.62	80.09	-
F Order, D**Global & Local Search End**	30.19%	68.47%	90.38%	3.17
Q Order, D**Global & Local Search End**	-15.29%	43.64%	78.47%	1.77
C Order, D**Global & Local Search End**	-89.73%	33.28%	86.92%	1.50

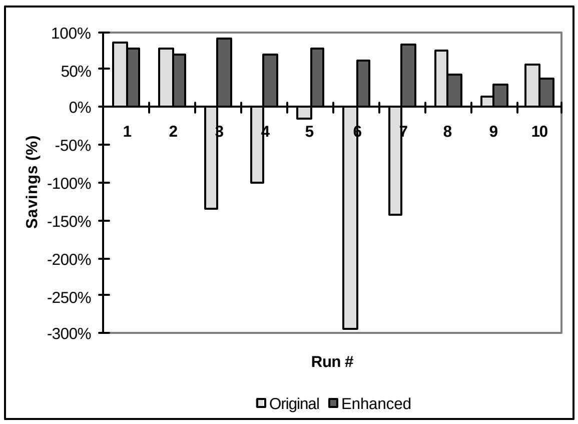

Figure 2. Comparison between Original and Enhanced SAHGA

Performance of Different Local Search Methods

In this section, we examine the effect of local search on the performance of SAHGA with fitness clustering and the other improvements.

The results for solving the test problems with the 5 algorithms for 10 runs with different random seeds are presented in Figure 3. The results show that the best performance is achieved by the uniform random walk algorithm (LS1), which is 30 to 90% faster than the SGA. The normal random walk algorithm (LS2) has almost the same reduction effect over the population but because the search is concentrated in a reduced part of the search space, the results are not as good. The third algorithm, (1+1)-ES, shows much better results than LS2, but worse than LS1. This shows that the adaptive features of LS3 were not helpful and a random search approach using LS1 is better. The next algorithm, random derivative, has the worst performance because it requires 2 function evaluations for each iteration, a cost that was not always worth the additional information gained. Finally, steepest descent is better than random derivative because the search is more effective. In any case, the performance is still worse than the performance of LS1 because of the cost of the additional function evaluations.

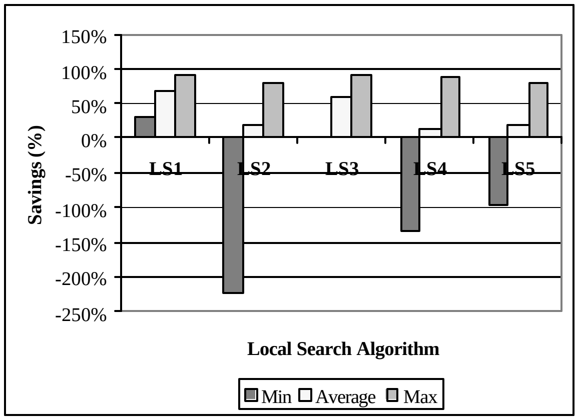

Figure 3. Results of Local Search Analysis

Conclusions and Recommendations

The results presented in this paper clearly indicate that a fitness-based clustering enhancement to the SAHGA algorithm, combined with improved stopping and local search criteria, enabled robust solution of a groundwater remediation design problem with 30 to 90% less effort than the SGA. The results of the local search analysis show that the best performance is attained for random walk with uniform distribution. This result is surprising given that this algorithm attains the least reduction in standard deviation in Fig. 1, which indicates that it had the least improvement in fitness per iteration. This finding results from both the low effort required for each computational iteration of the algorithm, compared with the derivative-based methods, and the apparent benefits of random search of the neighborhood surrounding the current point.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BES 97-34076 CAR and the U. S. Army Research Office under Grant No. DAAD 19-00-7-0389.

References

Espinoza, F., B. S. Minsker, and D. Goldberg. "Optimal Settings for a Hybrid Genetic Algorithm Applied to a Groundwater Remediation Problem". "*Bridging the Gap: Meeting the World's Water and Environmental Resources Challenges", Proceedings of the World Water and Environmental Resources Congress* (ISBN 0-7844-0569-7), Orlando, FL, ed. Don Phelps and Gerald Sehlke, American Society of Civil Engineers, Washington, DC. (2001). (http:/cee.ce.uiuc.edu/emsa)

Gopalakrishnan, G., Minsker, B. S., and Goldberg, D. E. "Optimal sampling in a noisy genetic algorithm for risk-based remediation design." *Journal of Hydroinformatics*. (2003)

Harik, G.R., Cantú-Paz, E., Goldberg, D. E., and Miller, B. L. "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)

Hogg, R., and Craig, A. *Introduction to mathematical statistics.* Macmillan Publishing Co., Inc., New York. (1978)

Iman, R. L., Davenport, J. M., and Zeigler, D. K. "Latin hypercube sampling (a programs user guide). Technical report SAND79-1473, Sandia Laboratories, Alburquerque. (1980)

Rechenberg, I. (1973). *Ecolutionsstrategie: Optimierung technischer systeme nach prinzipien der biologishen evolution.* Frommann-Iolzboog Verlag, Stuttgart, (1973)

Reed, P., Minsker, B. S., and Goldberg, D. E. "Designing a competent simple genetic algorithm for search and optimization", *Water Resources Research*, 36(12), 3757-3761. (2000)

Reed, P. Striking the balance: long-term groundwater monitoring design for multiple conflicting objectives. PhD dissertation, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign. (2002)

Schwefel, H. P. Evolutionsstrategie und numerische Optimierung. PhD dissertation, Department of Process Engineering, Technical University of Berin, Berlin, Germany. (1975)

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

Spitzer, F. *Principles of random walk.* D. Van Nostrand Company, Inc. (1964)