Efficient HPC Parallel Global Optimization Algorithms Applied to Nonlinear Partial Differential Eqn-Based Environmental Objectives

CHRISTINE ANN. SHOEMAKER (1)(2), MIN PANG (1),
(1): Civil and Environmental Engineering, Cornell University, Ithaca, NY14853, U.S.A
(2): Industrial System and Engineering, National University of Singapore, Singapore

ABSTRACT

We describe the application of a series of new and efficient synchronous and asynchronous parallel surrogate global optimization algorithms implemented in parallel on the Yellowstone supercomputer in USA with up to 128 processors. The application is to computationally complex and expensive partial differential equation models of flow and transport of chemicals in groundwater using data from the USA Environmental Protection Agency.

INTRODUCTION

Many real-world problems involve complicated simulation models. Examples such as watershed simulations, groundwater simulations may require extensive computational effort. For problems, such as identifying optimal decisions of environmental protection, designing optimal remediation strategies or improving the models by calibrating the model parameters, it is essential to use efficient optimization methods. Moreover, in many groundwater-related physical simulations, the relation between inputs and outputs are so complex that they can be viewed as black box functions. Optimization/simulation (O/S) strategy can be viewed as an efficient tool, in which simulators evaluate the inputs provided by optimizers, and then give simulation feedback to the optimizer to find better inputs.

In this paper, we will show that our parallel Surrogate-based global optimization algorithms are effective optimizers for complex, computationally expensive simulation models. On the one hand, it employs multiple processors in High Performance Computing (HPC) system to reduce computational and wallclock time; and on the other hand, it adopts the surrogate assisted strategy to guide the search and converges to the optimal result in relatively few iterations. In our studies, we implemented different Parallel Surrogated-based optimization algorithms, and applied them to various problems such as managing remediation in two polluted Groundwater superfund remediation sites in the U.S, Umatilla Chemical Depot and Blaine Naval Ammunition Depot. We also give an example for controlling land subsidence induced by groundwater extraction in Hang-Jia-Hu area in China and for calibrating for groundwater flow and transport models in Umatilla Chemical Depot site.

We used synchronous parallel algorithm Parallel Stochastic Radial Basis Function (Parallel RBF [1]) on groundwater remediation problem in Umatilla and Blaine sites. By comparing its performance with several alternative parallel algorithms, we established that Parallel Stochastic Radial Basis is an efficient algorithm to use on both the Umatilla and Blaine management problems. Parallel Radial Basis Surrogate with additional strategy for dealing with constraint expensive function (Parallel DYSOC) can solve problems with computational expensive constraints as in Hang-Jia-Hu land subsidence. In addition, we introduced asynchronous parallelism surrogate-based algorithms with early truncation strategy (SO-AET) that facilitate solving for computational expensive problems, which might cause unbalanced load from different trial inputs in different processors.

MODELS FORMULATION

Our parallel surrogate global optimization applications involve two groundwater remediation management problems, land subsidence controlling problems and models calibrations.

(1) Remediation Management Problem in Umatilla Chemical Depot

The formulation of remediation problem on Umatilla site consists of pumping system of eight pumping wells and two recharge basins. The modeling period is one management period of four years with pumping rate kept the same. The objective is to minimize the operation cost for this 4-year period constrained by the total pumping rate and cleanup level at the end of four years.

(2) Remediation Management Problem in Blaine Naval Ammunition Depot

For remediation problem on Blaine site, there is a similar objective goal, which is to minimize the cost for entire project duration. However, the modeling period of Blaine problem consists of 6 management periods of 5 years, and the system contains 15 pumping wells. The objective function is to minimize the management cost including fixed cost related to facility installation once per each management period and maintenance and operation cost for the whole project duration. There is a cleanup level constraint for each of the indicator containments at the end of the project.

(3) Land subsidence problem in Hang-Jia-Hu area

The goal of this study is to design groundwater exploitation management plans by controlling massive wells with constraints on induced subsidence. One of the formulations is to maximize total pumping rate under the cumulative land subsidence constraint at the end of management period.

(4) Calibration Problem in Umatilla Chemical Depot

We aim to identify the values for hydraulic conductivity zones in Umatilla flow and transport (i.e. MODFLOW and MT3D model) by minimizing the difference between observation data and the simulation data. Our observation data is assumed to contain weekly set of different observation wells for contaminant concentration during four years, and observation well heads at the end of year four.

ALGORITHMS AND SOFTWARE FOR MULTIMODAL OBJECTIVE FUNCTIONS

Surrogate-based Optimization (SO) algorithms (e.g. [3,8,9,10]) utilize surrogate models to guide the search for the optimal value of the objective function F(X) to reduce the number of expensive function evaluation necessary to find an accurate solution. This type of algorithms is especially suitable for global optimization problems with expensive function evaluations (e.g. minutes or longer). In addition, SO algorithms are derivative-free methods, which can solve a black-box problem efficiently. We use as a surrogate, a radial basis function (RBF), which interpolates all previously evaluated objective function values of F(X). In each iteration, we need to build an RBF model to approximate expensive function and use this model to select multiple points for simultaneous expensive function evaluation by multiple processors.

(1) Serial vs. Parallel

In the serial version, we update the RBF surface using only one newly evaluated point at each iteration. The optimization algorithm is changed to facilitate parallel computation so that in the parallel version, we chose to select different number of points evaluated based on number of processors used to test the performance of the algorithm. The parallelism scheme of the algorithm is in "master-slave" form. Each time after obtaining the points selected from candidate points using RBF surface and distance criteria, the "master" processor distributes P points to P "slave" processors to generate real function evaluation value. After all slave processors finished their job, the "master" processor gathers the results from P "slaves", then uses all real function evaluations values must update new RBF. The termination criterion is when the maximum evaluation is reached. By comparing the serial and parallel RBF surrogated based optimization algorithms, we can assess its efficiency utilizing the HPC system.

(2) Synchronous vs. Asynchronous

Synchronous parallelism means all parallel tasks are sent out at the same time and processors wait for all the tasks to be completed before doing more tasks. It is conventional method, but can result in idle time and load balancing problems. With asynchronous parallelism, the tasks are sent out as processors become available. Asynchronous is considerably more complex to code. We demonstrate a new algorithm SO-AET, an asynchronous parallel Surrogate-based optimization algorithm with early truncation strategy for objective function computation that has been designed for problems with characteristics are typical for goodness of fit objectives like sum of squared error. SO-AET has an advantage only if the algorithm permits asynchronous parallel computation since otherwise truncating the evaluation would just create a load-balancing problem. SO-SP is a synchronous parallel surrogate optimization method that would be the best choice if no asynchronous parallel platform is available. So, comparing SO-SP to SO-AT assesses the benefits of the asynchronous algorithm.

(3) Software

This analysis uses open source "pySOT" software ([3] Eriksson, Bindel, and Shoemaker, 2015), where SOT is an acronym for "Surrogate Optimization Toolbox". pySOT is flexible so a number of continuous algorithms (e.g. Regis & Shoemaker, 2007, 2009, 2013), integer algorithms (Mueller et al., 2013, 2014), and ensemble surrogate algorithms (Mueller and Shoemaker, 2014) can be implemented by appropriately choosing the pySOT options for each stage of the surrogate algorithm. pySOT is built on POAP in GitHub (Bindel et al. manuscript) with an asynchronous option.

The software is efficient for most multi modal simulation optimization problems and here we look at the important problem of parameter estimation for a groundwater model. The method is applicable to other geophysical problems (e.g. CO_2 sequestration (Espinet et al., 2013), and Climate Change (Mueller et al.2015))

IMPLEMENTATION

In this Simulation-Optimization problem, the simulation of groundwater flow is solved by MODFLOW with its latest version MODFLOW2005 [1], while the contaminant transport and fate is simulated by MT3DMS developed by Chunmiao Zheng [5].

MODFLOW is a three-dimension finite difference groundwater model maintained by U.S. Geological Survey. The input of MODFLOW is domain discretizing data, hydro-geological data including transitivity, initial head and hydraulic conductivity in each discretized node, and pumping data containing pumping well locations and pumping rate. The output from MODFLOW (i.e. the saturated thickness, fluxes across cell interfaces in all directions, and locations and flow rates of various sinks/sources, including transient groundwater storage) is used in MT3DMS to generate concentrations of the contaminants for the whole simulation period. Then, the concentrations are taken as indicator parameters of the optimizations to identity best pumping rates that minimize the management cost.

The implementation of simulation-optimization problem is run on the NSF Yellowstone supercomputer, a peta-scale computing resource in the NCAR-Wyoming Supercomputer Center. It is a 1.5-Petaflops-cluster computer, which contains 9,036 2.6-GHz Intel Xeon E5-2670 8-cores processors. This HPC system provides a large environment to test the performance of the algorithms, which we are using. The pySOT package has been installed at NSCC for analysis of Singapore surface water quality analysis, but the groundwater problems reported here are not yet running on NSCC in Singapore.

RESULTS AND DISCUSSION

(1) Umatilla remediation problems

We first applied serial Stochastic RBF algorithm on both the Umatilla and Blaine remediation problems using Yellowstone supercomputer described above. For Umatilla, we run 1000 expensive evaluations. Each simulation takes around 1.5 mins, so total wall clock time is 94314s (26.2 hr). Blaine is a more complicated problem; it takes 30 mins for each simulation, with a total of 8 days for 400 iterations.

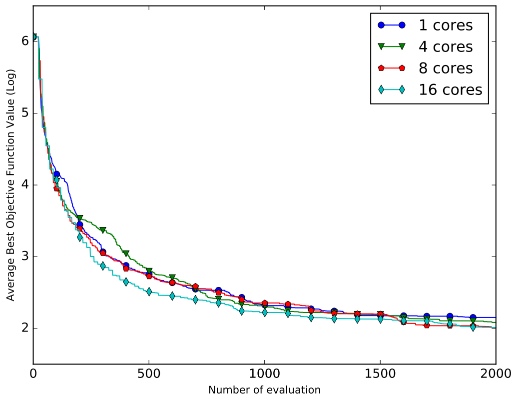

Figure 1 Progress Graph of Parallel Stochastic RBF on Umatilla Management Problem

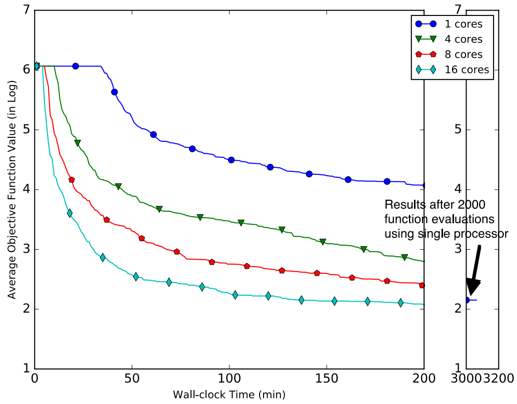

Figure 2 Average best results of Parallel Stochastic RBF against Wall clock time on Umatilla Management Problem

Comparing the results for Umatilla Problem using Parallel RBF and Serial RBF, we can see that we can obtain the best result sooner with more processors used as shown in Figure 1 and Figure 2. Both Figures show the results of best results averaged over 30 trials. And we can see that we can achieve better algorithm efficiency having a larger pool of working processors. In our study, we used up to 128 processors, and here we show the results up to 16 processors. In terms of parallelism efficiency, we compute the total wall clock time to reach to one accuracy level which is set as the averaged results obtained by using single processor after 2000 function evaluations for all four cases (with one, four, eight, sixteen processors). Then, the speed up is the ratio of the walk-clock time spent on parallel algorithm and the wall-clock time for the serial algorithm. Table 1 shows that we can get high efficiency more than 100% using up to 16 processors.

Table 1 Speed up and Efficiency of Parallel Stochastic RBF in Umatilla Remediation Management Problem

No. Procs	Speed up	Efficiency
4 proc	4.58	114%
8 proc	8.86	110%
16 proc	21.29	133%

(2) Umatilla calibration problem

We demonstrate a new algorithm SO-AET, an asynchronous parallel Surrogate-based optimization algorithm with early truncation strategy for objective function computation that has been designed for problems with characteristics are typical for goodness of fit objectives like sum of squared error.
"AET" stands for asynchronous parallelism with early truncation strategy, which helps to prematurely terminate the computationally expensive function. We compared the performance of SO-AET with synchronous Parallel Algorithms including Surrogate Optimization Algorithms in Synchronous Parallelism (SO-SP) and Shuffle Complex Evolutionary Algorithm (SCE-UA) as well as asynchronous algorithm which based on pattern search (APPSAPCK). SO-AET outperforms SO-SP, and SO-SP is the second-best algorithms, and both two algorithms outperform APPSPACK and SCE-UA.

CONCLUSION

Our parallel Surrogate based Optimization Algorithms in pySOT [3] show effective performance compared to its serial version as well as other alternative parallel algorithms. Both the surrogate surfaces and the parallelism paradigm help reduce computational budget, improve results quality and robustness in finding good optimal solution. In further research, the suggested methodologies could be extended for much larger scaled problems and to the problems from other application areas using traditional optimization methods.

Acknowledgements: This research funded by US NSF CISE directorate grant to Shoemaker at Cornell University and in Singapore by E2S2 CREATE grant and NUS start-up grant to Prof. Shoemaker.

REFERENCE

[1] Arlen W. Harbaugh, MODFLOW-2005, The U.S. Geological Survey Modular Ground-Water Model—the Ground-Water Flow Process

[2] Espinet, A., C.A. Shoemaker, C. Doughty, “"Estimation of Plume distribution for Carbon Sequestration using parameter estimation with limited monitoring data", Water Resources Research 49(7), 4442-4464, 2013

[3] Eriksson, D., D. Bindel, C.A. Shoemaker pySOT (see GITHUB) 2015 (over 16,000 downloads)

[4] Naval Facilities Engineering Command (NAVFAC) technical report TR-2237-ENV, 2004

[5] MT3DMS: A Modular Three-Dimensional Multispecies Transport Model for Simulation of Advection, Dispersion, and Chemical Reactions of Contaminants Groundwater Systems; Documentation and User’s Guide

[6] Mueller, J. and C.A. Shoemaker, “Influence of Ensemble Surrogate Models and Sampling Strategy on the Solution Quality of Algorithms for Computationally Expensive Black-box Global Optimization Problems, Journal of Global Optimization, 60 (2) p. 123-144, 2014

[7] Mueller, J. R. Paudel C.A. Shoemaker, J. Woodbury, Y. Wang, and N. Mahowald, CH4 parameter estimation in CLM4.5Bgc using surrogate global optimization. Geoscientific Model Development, 2015

[8] Regis, R and C. A. Shoemaker. “A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions”, INFORMS Journal on Computing, Vol. 19, No. 4, pp. 497-509, 2007

[9] Regis, R.G., C.A. Shoemaker, “Parallel Stochastic Global Optimization Using Radial Basis Functions,” INFORMS Jn. Of Computing 21(3), 411-426, 2009

[10] Regis, R, C.A. Shoemaker, “Combining Radial Basis Function Surrogates Dynamic Coordinate Search in High Dimensional Expensive Black-box Optimization”, Engineering Optimization 45 (5) 529-, 2013