SENSITIVITY ANALYSIS IN A PROBLEM OF REAXFF MOLECULAR-DYNAMIC FORCE FIELD OPTIMIZATION

K. S. Shefov [a], M. M. Stepanova [b]

St. Petersburg State University, 7-9 University emb, St. Petersburg, 199034, Russia

E-mail: [a] k.s.shefov@gmail.com, [b] mstep@mms.nw.ru

In a wide range of modern problems, it is required to estimate an influence of uncertainty of input parameters on uncertainty of an output value of a modeling function. In this contribution, we present algorithms for analyzing the sensitivity of a target function with respect to parameters in the problem of optimization of ReaxFF molecular-dynamic force field. In this particular case it allows one to effectively decrease the number of simultaneously optimized parameters. We compare the Sobol's global sensitivity indexes (SI) approach and the correlation analysis. Both methods are based on computations of the target function value on the set of pseudo- or quasi-randomly distributed points. The distribution derived is used for further computations of SI using Monte-Carlo technique and correlation coefficients. In the case of optimized ReaxFF force field one may spend up to several seconds to compute a value of the target function in a particular point. That is why it is important to perform calculations in parallel for multiple points. A parallel algorithm has been implemented in C++ using MPI. We compute Sobol's SI and coefficients of correlation of parameters variation and target function values variation while we optimize the force field for molecules and crystals of zinc hydroxide. We show that using of parameter set sorted by influence allows one to significantly increase convergence speed of the optimization algorithm and even completely exclude those parameters with relatively small influence.

Keywords: sensitivity analysis, reactive force field, molecular dynamics, parameter optimization, parallel algorithm, scalability

1. Introduction

In a wide range of modern problems, it is required to estimate the influence of the uncertainty of the input parameters on the uncertainty in the output value of the modeling function. In this paper there are presented algorithms of sensitivity analysis of an objective function to parameters in the problem of optimization of the molecular-dynamic force field ReaxFF. In this particular case, this effectively reduces the number of parameters simultaneously participating in the optimization. Two approaches are compared: one based on Sobol's global sensitivity indexes (SI) by and the other one uses correlation analysis. A parallel program is presented, its scalability is studied, and the calculation results for a particular task are given.

2. The Problem

A molecular-dynamic force field is characterized by a number of parameters: $U = f(p_1, p_2, \dots, p_N)$. These parameters are set before solving the Newton's equations of motion and do not alter during the simulation. Their number (N) may vary from 3 – 4 to several dozens. A procedure of a force field parameters search for a particular simulated system is referred to as a force field optimization. As a measure of optimality we use the parameter-dependent objective function (OF) $T(p_1, p_2, \dots, p_N)$. The OF is determined by the deviation of any characteristics of the simulated system obtained using the methods of MD, from those obtained by more accurate methods. The process of optimization is to search the force field parameters that bring the OF the minimal value.

As the MD force field $U(\vec{r})$ we use ReaxFF (Reactive Force Field) that is able to simulate chemical reactions [1]. To optimize the OF we use the multifactorial global search algorithm (MGSA) suggested by Strongin [2].

The MGSA is able to simultaneously optimize relatively small number of parameters (the optimal variant is 4). However, optimizing ReaxFF for a particular system may need the number of parameters to be of several dozens. Therefore the parameters being searched are divided into groups, e. g. of 4 pieces, which participate in the optimization. The question arises of how to sort the parameters by groups, and also how to exclude parameters that almost do not affect the change in the OF. This task could be solved with a help of sensitivity analysis. In this paper we consider the correlation analysis and the method of Sobol's sensitivity indices.

3. Correlation analysis

The objective function for searching MD force field parameters is a sum of terms depending on the parameters $p_1, p_2, \dots, p_N$. The procedure of correlation analysis requires the following steps.

1. Generate a sample of R (pseudo-)random sets of parameters $p_{1,i}, p_{2,i}, \dots, p_{N,i}$, $1 \le i \le R$, that are uniformly distributed over the search domain of the OF.

2. For each set of parameters compute the OF value and record the values of particular terms. One will obtain a $M \times R$ matrix of OF terms' values, where M is the number of terms of the OF. In total, one has two matrices: the $N \times R$ matrix of parameters P and the $M \times R$ matrix of OF terms F.

3. Normalize matrices P and F:

$$p_{ij} = \frac{p_{ij} - \overline{p_i}}{\sqrt{R}\sigma_{p,i}}, \qquad f_{ij} = \frac{f_{ij} - \overline{f_i}}{\sqrt{R}\sigma_{f,i}}.$$

Here $\overline{p_i}$, $\overline{f_i}$ are averages by columns, $\sigma_{p,i}$, $\sigma_{f,i}$ are dispersions by columns, R is the sample size.

4. Calculate the $M \times N$ matrix $C = P^T \cdot F$ of cross-correlations of parameters change and OF values change. The elements of the matrix, c_{ij}, are cross-correlation coefficients.

5. Sort the rows of the matrix C by the greatest absolute value of element in a row in descending order. Each row corresponds to a single parameter.

The correlation analysis will give good result only if the OF terms depend on the parameters monotonically, which is not always the case.

4. Sobol sensitivity analysis

A more efficient technique for estimating the influence of variables on a function is the calculation of global sensitivity indices (SI) suggested by I. Sobol and A. Saltelli [3]. The method of sensitivity indices, in contrast to the correlation analysis, does not require the presence of a monotonic dependence of the function on the variables.

Let $f(x)$ be a function of several variables $x = (x_1, x_2, \dots, x_n)$ defined and square integrable in the unit cube K_n. Then its ANOVA (Analysis of Variances) decomposition is

$$f(x) = f_0 + \sum_{s=1}^{n} \sum_{i_1<\dots<i_s} f_{i_1,\dots,i_s}(x_{i_1}, \dots, x_{i_s}) = \\ = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \dots + f_{1,2,\dots,n}(x_1, \dots, x_n), \qquad (1)$$

if $f_0 = \int_{K_n} f(x)dx$, and $\int_0^1 f_{i_1,\dots,i_s} dx_{i_p} = 0$, when $1 \le p \le s$. The inner sum in (1) is done by all the $i_1, \dots, i_s$, that satisfy inequalities $1 \le i_1 < \dots < i_n \le n$.

The decomposition (1) could be made with any complete orthonormal system of functions $\psi_0(x), \psi_1(x), \dots, \psi_k(x), \dots$, that includes the function $\psi_0(x) \equiv 1$.

The quantities $D_{i_1,\dots,i_s} = \int f^2_{i_1,\dots,i_s}(x_{i_1}, \dots, x_{i_s})dx_{i_1}, \dots, dx_{i_s}$ are called the dispersions. Here and below the sign $\int$ means integration from 0 to 1 by the corresponding variables. The quantity $D = \int f^2(x)dx - f_0^2$ is called the total dispersion. It is also true that. $D = \sum_{s=1}^{n} \sum_{i_1<\dots<i_s} f_{i_1,\dots,i_s}$.

The global sensitivity indices are the dispersion ratios

$$S_{i_1<\dots<i_s} = D_{i_1<\dots<i_s}/D.$$

In application single-dimension sensitivity indices S_i are most often used. With their help one is able to sort variables x_i: the bigger is S_i, the more influential is the variable x_i.

Let us consider an arbitrary group of variables $x_{k_1}, \dots, x_{k_m}$, where $1 \le k_1 < \dots < k_m \le n$, $1 \le m \le n - 1$. We shall denote them by a single letter $y = (x_{k_1}, \dots, x_{k_m})$; and let z be the aggregate of all the rest $n - m$ variables. Thus, $x = (y, z)$. With M let us denote the aggregate of indices $(k_1, \dots, k_m)$. For the set y let us introduce two types of global SI:

$$S_y = \sum S_{i_1,\dots,i_s}, \qquad S_y^{\text{tot}} = \sum S_{i_1,\dots,i_s}.$$

In S_y the summation is produced for all the groups of $i_1, \dots, i_s$, where all the $i_p \in M$. In S_y^{tot} the summation is produced for all the groups of $i_1, \dots, i_s$ so that at least one index $i_p \in M$.

One is able to calculate the SI S_y and S_y^{tot} by integrating using (pseudo/quasi) Monte Carlo technique. Let x and x' be the points of K_n and let $x = (y, z)$, and $x' = (y', z')$. In the paper [3] it is shown that

$$D_y = \int f(x)f(y, z')dxdz' - f_0^2,$$

$$D_y^{\text{tot}} = \frac{1}{2}\int [f(x) - f(y', z)]^2 dxdy' = D + f_0^2 - \int f(x')f(y, z')dx'dy.$$

In order to compute all the one-dimensional ($y = (x_i)$) SI S_y and S_y^{tot}, in each sample of the Monte Carlo method one needs to use two n-dimensional random points x and x' and calculate the value of the function $n + 2$ times: $f(x)$, $f(x')$, and $\omega_i = f(x'_1, \dots, x'_{n-1}, x_n, x'_{n+1}, \dots, x'_n)$, when $1 \le i \le n$.

5. LPτ Sequences

One may select the points of parameters in a pseudo-random way, but it will not provide a uniform cover of the whole unit cube K_n. The alternative is to use LPτ sequences [4]. These strictly ordered sequences are based on multiple binary division of the domain by all the dimensions. They allow one to construct a mesh on a hypercube so that the mesh nodes fill it as uniformly as possible.

Let us compare a mesh of 16 points of this sequence with the simple cubic mesh for the square (Figure 1). The points of the simple mesh allow one to calculate function values only in 4 different values of each of the two variables. The mesh built by LPτ allows one to obtain function values in 16 different values of each variable for the same total number of points. Thus, LPτ provides more efficient distribution of points.

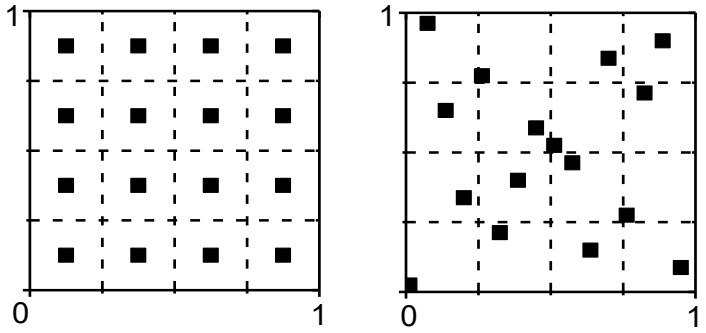

Figure 1. Simple mesh (left) and LPτ mesh (right)

6. Implementation and Scalability

We implemented parallel programs for sensitivity analysis in C++ using MPI technology. The use of sensitivity analysis in the problem of ReaxFF force field optimization has an important feature that the time of calculation of a single OF value significantly exceeds the time of all other operations, i. e. data exchange between processes of interaction with the file system. OF values are computed with the help of LAMMPS molecular dynamic simulation library [5]: it calculates values of energies and forces for several dozens of molecules.

In Figure 2 the general scheme of the implemented programs is presented. In both cases (correlation analysis and sensitivity indices approach) the scheme is the same. At first one generates an array of N-dimensional points (ReaxFF parameters). The array is divided into equal parts between all the parallel processes and is sent between them to compute the OF values.

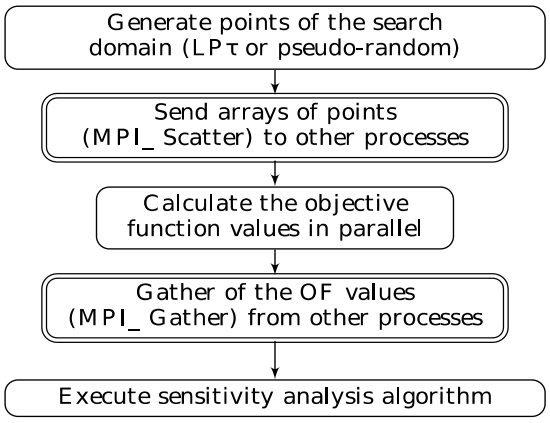

Figure 2. General scheme of programs

In the stage of OF values calculation the data exchange is not needed. Each process computes the values independently. The calculated OF values are gathered by the main process which computes correlation coefficients and Sobol's SI. Due to no data exchange in the main loop of the programs good scalability is expected. This was verified on a cluster of the following configuration.

14 nodes: 2 × 4 core CPU Intel® Xeon® E5335 2.00 GHz, 16 GB RAM; 112 cores in total;
OS: CentOS Linux 7 (Core); MPICH2 v1.4.1p1.
Hardware is provided by Resource Center Computer Center of St. Petersburg State University.

We measured the dependence of acceleration of programs' time of operation on the number of CPU cores in use for 1, 7, 14, 28, 56, and 112 pieces. The computations we done for 112 points of the search domain; when using all the 112 cores, we had one point per process. The average time of computation of a single OF value by a single process is 12.4 seconds.

Figure 3 shows the measured dependence for both programs (CC and SI), since the values of acceleration match with high accuracy. Linear approximation is done with the least squares technique. The efficiency of parallelism (the ratio of acceleration increase and cores number increase) is (71.2±1.3) %.

The time of operation of the serial part of the programs, i. e. the time of SI or CC computation is less than 1 second for 112 points and is about 30 seconds for 130000 points, which is significantly smaller than the time of the parallel part of the programs.

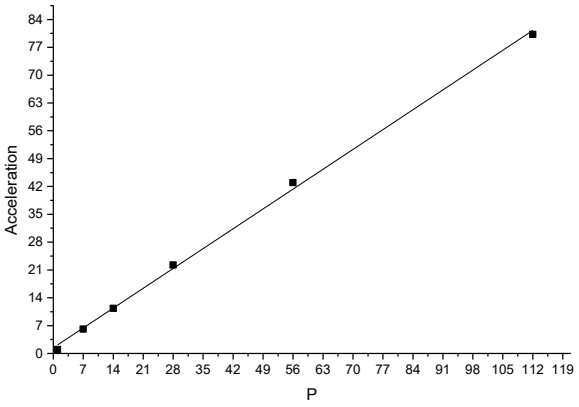

Figure 3. Accel. dependence on the cores number

7. Application

The programs presented have been applied to sensitivity analysis of the objective function to the parameters of ReaxFF force field in its optimization for Zn – O – H compounds.

The sensitivity indices (SI) and correlation coefficients (CC) calculation is performed for 58 parameters on 130000 points. The diagram on Figure 4 depicts relative values of SI and CC for 24 parameters of ReaxFF. For the unity we take the greatest SI in the first case, and the greatest CC in the second case. The parameters on the diagram are sorted by SI in descending order. On the chart for CC white hatching highlights those parameters that were not included in the first 24 in the correlation analysis. The number of such parameters is not large relative to the length of the whole list. One has to note that 5 parameters with the greatest CC are also located in the beginning of the list by SI. In general, the order of the parameters for the SI and CC is noticeably different.

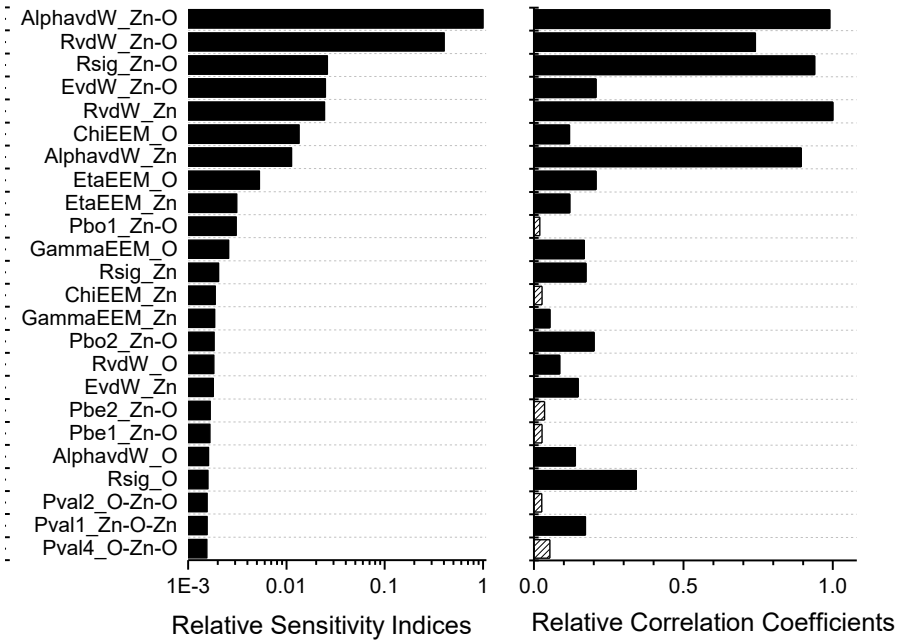

Figure 4. SI and CC for ReaxFF force field parameters for ZnO

Using MGSA method we perform ReaxFF optimization by 24 parameters: first with the greatest CC, then with the greatest SI. The optimization is done in groups of 4 parameters. The sequence of the groups is defined by sorting the parameters by CC and SI, respectively. The loop of groups is repeated until the difference of OF values of the two consecutive iterations becomes less than a predefined tolerance. If the order of the groups in the optimization is defined by CC, the algorithm requires 2.5 loop passes to converge. When the order is defined by SIs, the algorithm needs

1.5 passes. Thus, sorting the parameters using the Sobol's SI gives for this problem a better result than using the correlation analysis.

8. Conclusion

Sensitivity analysis is useful in the problem of optimizing of a large number of parameters of ReaxFF force field in groups. The approach allows one to exclude those parameters that have little influence on the objective function and also effectively sort them by influence.

The developed parallel program for Sobol sensitivity indices and correlation coefficients calculation is effectively scalable on a computational cluster.

To sort the parameters in ReaxFF optimization for zinc — oxygen — hydrogen compounds it is better to use Sobol sensitivity analysis than the correlation analysis, since the procedure converges faster in the first case.

References

[1] K. Nomura, R. K. Kalia, A. Nakano, P. Vashishta, J. L. Landa. A scalable parallel algorithm for large-scale reactive force-field molecular dynamics simulation // Comp. Phys. Comm. — 2008. — Vol. 178(2). — P. 73 – 87.

[2] Stepanova M.M., Shefov K.S., Slavyanov S.Yu. Multifactorial global search algorithm in the problem of optimizing a reactive force field // Theoretical and Mathematical Physics. — 2016. — Vol. 187, Issue 1. — P. 603 – 617.

[3] Sobol' I. M, Global'nye pokazateli chuvstvitelnosti dlya izucheniya nelineynykh matematicheskikh modeley [Global sensitivity indices for exploration of non-linear mathematical models] // Matematicheskoye modelirovanie [Mathematical modeling] — 2005. — Vol. 17 (9). — P. 43 – 52. (in Russian)

[4] Sobol' I. M., Statnikov R. B. Vybor optimal'nykh parametrov v zadachakh so mnogimi kriteriyami [Choise of optimal parameters in problems with multiple criteria]. — M.: «Drofa». 2006. — 175 p. (in Russian)

[5] LAMMPS package: https://lammps.sandia.gov/ (accessed 30.09.2018)