Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Data Reduction Through Increased Data Utilization in Chemical Dynamics Simulations

Misha Ahmadian [a], Yu Zhuang [a], William L. Hase [b], Yong Chen [a]

[a] *Texas Tech University, Department of Computer Science, Lubbock, TX, 79409-3104, United States*
[b] *Texas Tech University, Department of Chemistry and Biochemistry, Lubbock, TX, 79409-1061, United States*

ARTICLE INFO

Article history:
Received 25 May 2016
Received in revised form 11 November 2016
Accepted 8 June 2017
Available online xxxx

Keywords:
Data utilization
Machine learning
Approximation
Interpolation
Chemical dynamic simulation
VENUS

ABSTRACT

Many scientific applications consist of heavy computational and analysis workload on data, and often require producing intermediate data for ongoing calculations. For instance, chemical dynamics simulations are known as heavy workload applications in terms of calculation in many aspects. There is a strong desire of seeking a solution to minimize expensive calculations by replacing them with light-weight ones. VENUS is one of these chemical dynamic simulation software packages known as classical chemical dynamics simulation, with scalar executing code and heavy calculation process. In this research, we introduce an innovative approximation method by storing, managing, and leveraging intermediate data (results) in order to speed up expensive calculations. The key idea is a newly introduced data interpolation method that leverages data points from previous calculations. The newly proposed method is a general approach that can be applied to a variety of scientific applications and disciplines. In this research, we focus on chemical dynamics simulations and the VENUS code and have developed a prototype of the data interpolation method for reduced computations. The proposed computation reduction method through increased data re-use can increase the efficiency and productivity of scientific simulations, thus can have an impact on scientific discovery powered by high performance computing simulations.

1. Introduction

Scientific applications increasingly utilize large-scale data in various fields including physics, astrophysics, climate studies, bioinformatics [1], and chemistry. In many of these science and engineering investigations, there are data that are of critical importance but highly expensive to generate, e.g. obtainable through time-consuming experimental or computational processes. Examples of highly expensive data include those produced by heavy scientific calculations, which might take hours, days, or even weeks to compute. If such critical and expensive datasets are also very large, there is an incentive to reduce the amount of such data, since reduction of expensively computed data also means reduction of computation time. However, what is critical is that the reduction of data should not lead to, within a problem-dependent threshold, the loss of the information that is carried by the original larger set of data. Thus, for scientific applications with large computationally expensive datasets, the objective of this research is to develop a computationally efficient data modeling procedure to replace the computation of some, hopefully as many as possible, data computed through highly expensive processes, and this data modeling procedure must be accurate so that the model data carries almost the same scientific information as the original expensive data.

In this research, we concentrate our approach of replacing expensive data computation by an efficient procedure of data modeling in chemical dynamics simulations, where the potential energy needed in every time step of a simulation has to be computed using a highly expensive computation procedure. Potential energy computation is the most expensive part in each step of a chemical dynamics simulation, whether using empirical formulas or using quantum mechanical electronic structure theories. Parallelization of potential energy calculation is a widely adopted approach to reducing the computation time, as done in many electronic structure calculation packages including NWChem [2,3] and in analytic empirical formulas [4]. Potential energy data calculated based on a quantum mechanical electronic structure theory are called *ab initio* potential energy data and they are far more computation intensive than empirical formula-based calculations. In this paper, the ab initio data are our targets to be replaced by modeled data at

E-mail addresses: misha.ahmadian@ttu.edu (M. Ahmadian), yu.zhuang@ttu.edu (Y. Zhuang), bill.hase@ttu.edu (W.L. Hase), yong.chen@ttu.edu (Y. Chen).

http://dx.doi.org/10.1016/j.bdr.2017.06.005
2214-5796/© 2017 Published by Elsevier Inc.

as many time steps as possible while trying to maintain a desired simulation accuracy.

While data modeling methods are problem dependent and require specific domain science knowledge to design an effective one, this research focuses on machine learning techniques for problems where the originally expensive data have a broadly existent property. Using our techniques, we have developed a method for modeling ab initio potential energy data, and implemented our data modeling method in VENUS, a chemical dynamics simulation software package. VENUS [2,5,6] has its own in-package suit of analytic potential energy models and is also linked with several electronic structure calculation packages, including NWChem [2,3], for generating accurate ab initio potential energy data. In the implementation, our model data replace the ab initio data generated by NWChem.

The rest of this paper is organized as follows. We will explain the chemical dynamics simulations and the VENUS simulation software in Section 2. The descriptions of our data modeling technique and methodology are given in Section 3. The method of using modeling data to replace expensively generated data, including its application to chemical dynamics simulations, is described in Section 3. Section 4 presents experimental results, and Section 5 concludes this study.

2. Classical trajectory chemical dynamics simulation

Classical trajectory chemical dynamics simulation provides a useful and generally applicable investigation tool for dynamics studies including gas–surface collisions [12], energy transfer and chemical reaction in gas-phase [13], intramolecular vibrational energy distribution [7], unimolecular decomposition and conformational change [14,15], and, intramolecular energy transfer and chemical reaction [16,17]. For these calculations, the potential energy function V, the potential gradient dv/dq, and in some cases the Hessian H, are required in the process of calculating an ensemble of trajectories, and each trajectory will be determined by numerically integrating the classical equations of motions [2].

A general and accurate approach used in chemical dynamics simulation is to calculate the potential energy data directly from electronic structure theory. VENUS contains a set of analytic potential energy functions and is also integrated with electronic structure calculation packages such as NWChem [3], MOLPRO [18], GAMASS [19], etc.

In chemical dynamics simulations, initial conditions of the reactants for chemical reactions are given for calculating an ensemble of trajectories. Each trajectory is evaluated by numerically integrating either Hamilton's or Newton's equations of motion. In an ab initio chemical dynamics simulation, the potential energy data including the energy gradients are calculated by an electronic structure program (e.g. NWChem).

VENUS software package [2,6] is a general Monte Carlo classical trajectory program, when calculating a classical trajectory, the execution of the program begins by reading Cartesian coordinates q and moments of inertia p. The selection of initial conditions calls a subroutine that integrates classical equations of motion, using the potential energy V, and its gradient dv/dq and Hessian H to produce dynamics results including the vibrational energy within the molecule. At the final stage of program, cross sections, scattering angels, product energy distribution, rate constant, etc. will be analyzed. The flow of VENUS process is shown in Fig. 2.1 [6].

Based on physical systems of interest in molecular dynamics simulations, VENUS obtains ab initio potential energy data from electronic structure calculation software, like NWChem, and it also provides a variety of analytic potential energy functions to build blocks of potential surfaces, and molecular model systems.

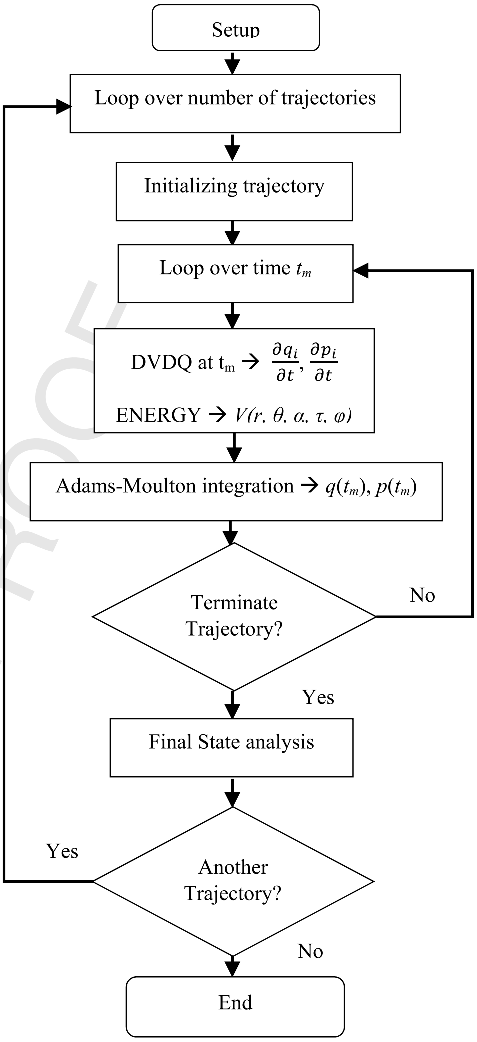
Setup
Loop over number of trajectories
Initializing trajectory
Loop over time t_m
DVDQ at t_m → $\frac{\partial q_i}{\partial t}, \frac{\partial p_i}{\partial t}$
ENERGY → $V(r, \theta, \alpha, \tau, \varphi)$
Adams-Moulton integration → $q(t_m), p(t_m)$
Terminate Trajectory?
No
Yes
Final State analysis
Yes
Another Trajectory?
No
End

Fig. 2.1. Flowchart of the VENUS program.

3. Modeling-enabled data reduction through increased data utilization

3.1. The general framework

Data reduction in this work refers to the reduction of data generated by expensive process of calculations, not covering general data reduction techniques like de-duplication, compressions, dimension reduction. We focus on expensive data since, in many cases, inexpensive data can be re-generated with very low costs.

Our strategy for reducing the costly computations for data generation is to replace some of them by data modeling. While data modeling methods do need domain specific knowledge to achieve high modeling accuracy, there is a feature that exist widely in scientific data. Many data generated in scientific or engineering processes consist of data from controllable parameters of the scientific or engineering processes, called input parameter data, and other data resulting from processes with the controllable parameters as input parameter data. In many cases, the resulting data are continuous functions of the input parameter data. In another word, the causal relationship between the input data and the resulting data may have a relationship that is of a continuous function. Even for discrete data in input–output pairs, relations resembling continuous functions also exist. For instance, many discrete optimization methods, e.g. the simulated annealing and the genetic algorithm, are based on the assumption that close input data will, in high

Please cite this article in press as: M. Ahmadian et al., Data Reduction Through Increased Data Utilization in Chemical Dynamics Simulations, Big Data Res. (2017), http://dx.doi.org/10.1016/j.bdr.2017.06.005