Commun. Comput. Phys.
doi: 10.4208/cicp.110613.280214a

Vol. **16**, No. 2, pp. 467-490
August 2014

On Stochastic Error and Computational Efficiency of the Markov Chain Monte Carlo Method

Jun Li[1,*], Philippe Vignal[1,2], Shuyu Sun[3] and Victor M. Calo[1,3]

[1] *Center for Numerical Porous Media, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.*
[2] *Material Science and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.*
[3] *Applied Mathematics and Computational Science, Earth Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.*

Received 11 June 2013; Accepted (in revised version) 28 February 2014

Available online 12 June 2014

Abstract. In Markov Chain Monte Carlo (MCMC) simulations, thermal equilibria quantities are estimated by ensemble average over a sample set containing a large number of correlated samples. These samples are selected in accordance with the probability distribution function, known from the partition function of equilibrium state. As the stochastic error of the simulation results is significant, it is desirable to understand the variance of the estimation by ensemble average, which depends on the sample size (i.e., the total number of samples in the set) and the sampling interval (i.e., cycle number between two consecutive samples). Although large sample sizes reduce the variance, they increase the computational cost of the simulation. For a given CPU time, the sample size can be reduced greatly by increasing the sampling interval, while having the corresponding increase in variance be negligible if the original sampling interval is very small. In this work, we report a few general rules that relate the variance with the sample size and the sampling interval. These results are observed and confirmed numerically. These variance rules are derived for the MCMC method but are also valid for the correlated samples obtained using other Monte Carlo methods. The main contribution of this work includes the theoretical proof of these numerical observations and the set of assumptions that lead to them.

AMS subject classifications: 76T99, 82B05, 82B80, 62M05, 65C40

Key words: Phase coexistence, Gibbs ensemble, molecular simulation, Markov Chain Monte Carlo method, variance estimation, blocking method.

*Corresponding author. *Email addresses:* lijun04@gmail.com (J. Li), philippe.vignal@kaust.edu.sa (P. Vignal), shuyu.sun@kaust.edu.sa (S. Sun), victor.calo@kaust.edu.sa (V. M. Calo)

1 Introduction

The Monte Carlo method has successfully been applied to a wide variety of applications, which include the solution of integral equations by the Markov Chain Monte Carlo (MCMC) method [1], the Boltzmann equation by the Direct Simulation Monte Carlo (DSMC) method [2] and stochastic partial differential equations by a multilevel Monte Carlo method [3]. We focus our discussion on the MCMC method. An essential part of many scientific problems is to evaluate an integral in a high-dimensional space $\vec{X}$ with the integrand containing a weighting function $f(\vec{X})$ (probability distribution function of the configuration $\vec{X}$) which is large in some area but close to zero almost everywhere else. The computational cost of evaluating the integral by conventional quadrature schemes is prohibitive since it demands a large number of quadrature points inside a high-dimensional space. This integral can be estimated by the average value of the integrand over a large number of configurations sampled inside the domain randomly, independently and uniformly, using the Monte Carlo (MC) method. Metropolis and Ulam [4] (see [5]) dubbed this simulation method *Monte Carlo* since it uses a large number of random fractions generated by a computer. The accuracy of the MC method can be improved by using the importance sampling scheme [6], which generates configurations non-uniformly but according to an artificially selected probability density function $g(\vec{X})$, which is close to $f(\vec{X})$, so that more probability mass is assigned to those configurations with higher probability [5–7]. In order to ensure the sampled configurations remain independent, the process demands the primitive function $G(\vec{X})$ of $g(\vec{X})$ and its inverse function $\vec{X}(G)$. Unfortunately, it is not feasible to find such $g(\vec{X})$ in most applications of interest. Rather than generating independent configurations, the Metropolis method [1], which still uses the importance sampling idea, generates (possibly) correlated configurations from the original $f(\vec{X})$ by a Markov chain. The Markov chain makes the algorithm simple and universal. This method is known as MCMC method [7]. Since the samples are correlated with each other, the variance of MCMC simulations with the same sample size is larger than the variance of the MC simulations using independent configurations. Additionally, the variance of MCMC simulations usually depends on the sampling interval.

The use of averages is common in scientific studies and many quantities related to thermal equilibria are averaged properties, measured in real experiments over large numbers of particles and long time intervals. If the ergodic hypothesis applies to the system at the molecular level [5], we can compute those quantities by ensemble averaging instead of time averaging using the probability distribution function $f(\vec{X})$, known from the partition function of the equilibrium state, an idea stemming from statistical mechanics. The MCMC method is a powerful tool based on ensemble averaging idea that can be used to calculate the quantities related to the thermal equilibrium state.

A system with fixed particle number N, volume V, and temperature T can be described by a canonical ensemble (constant-NVT), with the probability distribution function containing only the coordinates of the N particles as independent variables. This description is valid for systems where the quantities of interest only depend explicitly