Computer Physics Communications 196 (2015) 143–148

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

# Strategies for non-uniform sampling of molecular dynamics phase space trajectories of relaxation phenomena

Matthew P. Kroonblawd, Thomas D. Sewell *

*Department of Chemistry, University of Missouri-Columbia, Columbia, MO 65211-7600, USA*

ARTICLE INFO

*Article history:*
Received 17 March 2015
Accepted 30 May 2015
Available online 6 June 2015

*Keywords:*
Time-series analysis
Non-uniform sampling
Molecular dynamics
Relaxation phenomena
Shock waves
Particle methods

ABSTRACT

Non-uniform phase space trajectory sampling schedules were developed for time-domain analysis of molecular dynamics simulations of transient processes to reduce the data storage costs to record the trajectory for post-processing and permanent archival. Sampling periods are defined as functions of time such that the spacing between sampled points is either exactly or approximately constant in the logarithmic-time domain. A non-uniform sampling schedule was applied to uniformly sampled time histories of translational kinetic energy for translationally excited molecules in a crystal. Direct comparison of the non-uniformly sampled time history to the uniform one reveals that the non-uniform time history accurately captures the relaxation process with a 93% reduction in data storage cost for the conditions studied. Implications for frequency-domain analysis are discussed. A sampling schedule for simulations of supported shock waves is presented that provides an approximate 60% reduction in data storage cost.



## 1. Introduction

Non-uniform data sampling has been studied and applied in a wide variety of fields including nuclear magnetic resonance spectroscopy [1,2], optics [3], holography [4], climatology [5], planetary mapping [6], low-power systems design [7], and speech coding [8]. A primary reason to use non-uniform sampling is to reduce the time required for data acquisition or the data storage cost. Data storage costs to keep phase space trajectories (hereafter *trajectories*) from molecular dynamics (MD) and other particle-based simulations for post-processing can be prohibitively expensive, especially if many particles $N$ or independent trajectories are considered. If a large ensemble of trajectories is considered, the trajectories may not even be saved at all. Consequently, time histories of simulation metrics (time-domain signals), such as kinetic and potential energy components, the stress tensor, or strain, are routinely computed from the trajectory on-the-fly during integration and are recorded more frequently than the underlying trajectory. In a recent trillion-atom MD simulation, for which the data storage requirement for a single frame with only single-precision positions was 12 TB, even trajectory snapshots were rendered on-the-fly [9]. Inability to save adequately sampled trajectories due to limited data storage capacity places an enormous constraint on post-processing analysis. These constraints will likely only increase in the future as increases in CPU power and massively parallel computing extend the spatial and time scales accessible to all-atom MD.

Molecular dynamics is a particularly useful tool for the study of relaxation processes in materials on sub-micron and sub-microsecond scales [9–19]. In many cases, a direct perturbation is applied to a selected subset of system degrees of freedom and the response of the system is calculated. The relaxation may exhibit multiple characteristics corresponding to multiple processes operative on disparate time scales. For example, time-domain signals from a relaxation process may exhibit initial short-lifetime characteristics, such as a damped oscillation, that differ considerably from the long-lifetime characteristics, such as the decaying exponential indicative of a (pseudo-)first-order process or a stretched exponential [20] as is used in some phenomenological relaxation models. That is, the sampling period (the interval between sampled time frames) required to accurately capture the relaxation phenomenon changes in time. Reducing the data storage cost to save adequately sampled trajectories facilitates *a posteriori* time-domain analysis by easing the constraint to predict what metrics will be of interest for computation on-the-fly during trajectory integration. In this brief report we define and implement a non-uniform trajectory-sampling schedule for which the sampling period is a function of time. We demonstrate for a particular case of interest that the non-uniform sampling

* Corresponding author.
*E-mail address:* SewellT@missouri.edu (T.D. Sewell).

http://dx.doi.org/10.1016/j.cpc.2015.05.026
0010-4655/

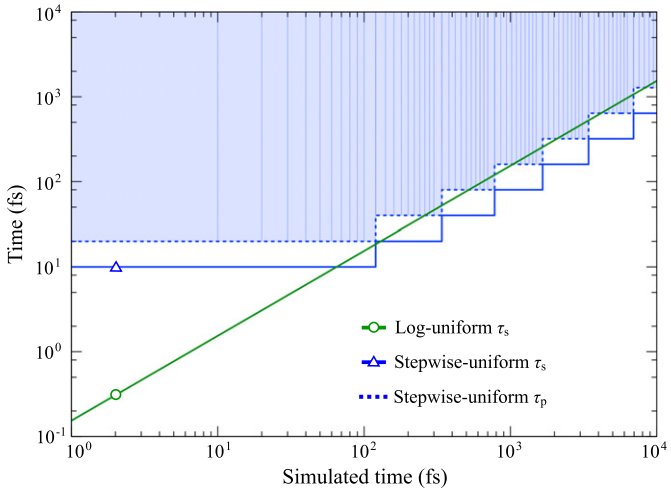


**Fig. 1.** Comparison of a base-10 logarithmic-uniform sampling period $\tau_s$ (solid green curve with circle) as defined in Eq. (5) for the particular case $M = 65$, $t_1 = 1$ fs, and $t_M = 9500$ fs, to the stepwise-uniform sampling period (solid blue curve with triangle) as defined in Eq. (7) with initial time $t_1 = 10$ fs, base period of $\tau_b = 10$ fs, and $Q = 10$. Note that both axes are plotted on a base-10 logarithmic scale. The dashed blue curve is the minimum process lifetime $\tau_p$ corresponding to the stepwise-uniform sampling schedule as determined by the Nyquist criterion given by Eq. (1) and the blue-shaded region above $\tau_p$ indicates the range of lifetimes of processes that can be accurately sampled. The vertical lines in the shaded region indicate times at which the trajectory is sampled. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

schedule reproduces the key time-domain features of a signal from a relaxation process with 93% fewer time frames as compared to a uniform sampling schedule applied across the same overall simulated time interval. Considerations for optimally sampling the frequency response and extensions for non-uniform sampling of a system with a moving perturbation such as a supported shock wave are also discussed.

## 2. Non-uniform sampling methods

One approach to sampling complicated processes in an MD simulation involves setting the sampling period based on the Nyquist–Shannon sampling theorem [21]; that is,

$$\tau_s = \frac{1}{2}\tau_p, \tag{1}$$

where $\tau_s$ is the sampling period (the Nyquist interval) and $\tau_p$ is the process lifetime. (A glossary of mathematical symbols is given in Appendix A.) Making $\tau_s$ short enough to accurately characterize the initial short-lifetime features in a signal corresponds to using a time resolution that is many times finer (and proportionally more resource intensive) than is necessary to characterize the long-lifetime features.

Because time-domain signals with both short- and long-lifetime features are commonly plotted against the logarithm of time [11,12,17], a reasonable approach to minimizing the data storage cost while retaining the essential features of the signal is to sample the trajectory at times that are uniformly distributed in the $\log_B(t)$ domain. The most common bases one would likely use are $B = 2$, e, and 10. For an ordered sequence of sampled time frames $t_j = \{t_1, t_2, t_3, \ldots\}$, where $t_1 > 0$, a given frame $t_j$ at which the trajectory is sampled using a logarithmic-uniform schedule is defined recursively in terms of a particular $t_1$ as

$$\log_B(t_j) \equiv \log_B(t_{j-1}) + \Delta_B, \tag{2}$$

where $\Delta_B$ is the constant separation between the sampled frames in the $\log_B(t)$ domain. The closed-form expression for the specific time frames is

$$t_j = t_1 B^{(j-1)\Delta_B}. \tag{3}$$

The time-dependent sampling period is defined as

$$\tau_s(t_j) \equiv t_j - t_{j-1}, \tag{4}$$

which for the logarithmic-uniform sampling schedule yields

$$\tau_s(t_j) = t_1 B^{(j-1)\Delta_B}[1 - B^{-\Delta_B}]. \tag{5}$$

A reasonable approach to determine the constant $\Delta_B$ is to fix a chosen number of sampled time frames $M$ corresponding to some chosen time $t_M$, for instance recording $M = 65$ frames between $t_1 = 1$ fs and $t_M = 9500$ fs (as will be used later in an example application). With this approach the separation constant is given by

$$\Delta_B = \frac{\log_B(t_M) - \log_B(t_1)}{M - 1}. \tag{6}$$

We observe that some reports [11,12,17] include figures that depict seemingly non-uniformly sampled data, but those reports either omit definitions and discussion of sampling schedules and their consequences [11,17] or are verified to have used uniform sampling [12].

In practice, rather than rigorously apply a logarithmic-uniform sampling schedule, it is easier to approximate $\tau_s(t_j)$ in a stepwise-uniform manner wherein each plateau is an integer multiple of a base period $\tau_b$. We make the distinction that only linear-uniform schedules (*i.e.* those with a static $\tau_s$) are uniform in the context of signal analysis; a stepwise-uniform sampling schedule is a non-uniform schedule even though it is composed of piecewise linear-uniform segments. Fig. 1 shows a particularly advantageous choice for a stepwise-uniform sampling schedule in which the sampling period is doubled every $Q$ frames,

$$\tau_s(t_j) = \tau_b 2^{\left\lfloor \frac{j}{Q} \right\rfloor}, \tag{7}$$

where $Q$ is a constant positive integer ($Q \in \mathbb{Z}^+$) and $\lfloor x \rfloor$ is the largest integer not greater than $x$ (the floor function). A closed-form expression for the sampled times $t_j$,

$$t_j = t_1 + \tau_b \sum_{k=2}^{j} 2^{\left\lfloor \frac{k}{Q} \right\rfloor}, \tag{8}$$

is determined by Eq. (4), $\tau_b$, and the initial time $t_1$. One can determine reasonable values for $\tau_b$ and $Q$ given the period of the highest-relevant-frequency mode in the system and an order-of-magnitude estimate for the overall relaxation lifetime. Although