Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

MARCUS WEBER

An efficient analysis of rare events in canonical ensemble dynamics

ZIB-Report 08-36 (January 2009)

An efficient analysis of rare events in canonical ensemble dynamics

Marcus Weber

Zuse-Institute Berlin, Takustraße 7, D-14195 Berlin, Germany

Abstract. For an analysis of a molecular system from a computational statistical thermodynamics point of view, extensive molecular dynamics simulations are very inefficient. During this procedure, at lot of redundant data is generated. Whereas the algorithms spend most of the computing time for a sampling of configurations within the basins of the potential energy landscape of the molecular system, the important information about the long-time behaviour of the molecules is given by transition regions and barriers between the basins, which are sampled rarely only. Thinking of molecular dynamics trajectories, researchers try to figure out which kind of dynamical model is suitable for an efficient simulation. This article suggests to change the point of view from extensive simulation of molecular dynamics trajectories to more efficient sampling strategies of the conformation dynamics approach.

Keywords: metastability, conformation dynamics, rate matrix, molecular dynamics
PACS: 02.70.Lq, 02.70.Hm

INTRODUCTION

Classical molecular dynamics simulation (MD) is a widely used method for the analysis of molecular interactions. The equations of motion are solved on the basis of a given force-field which models the different mechanical aspects of covalent bonds and non-covalent interactions. Researchers try to estimate the essential behaviour of a molecular system by a statistical evaluation of the generated states. A state $x = (q, p)$ of a molecular system consisting of n atoms is given by a $6n$-dimensional vector of $3n$ position coordinates $q \in \Omega$ and $3n$ momentum coordinates $p \in \Gamma$. The total energy $H(q, p)$ of a state is the sum of the kinetic energy $K(p)$ (only depending on the momentum coordinates) and the potential energy $V(q)$ (only depending on the position coordinates). Thus, H is separable. In MD, the equations of motion are solved:

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial q}. \end{aligned} \tag{1}$$

Standard MD simulation, however, would lead to incorrect statistical results if one wants to analyze a system at a given constant temperature T, with constant volume v and constant number of particles n. For this nvT-ensemble, Boltzmann derived the theoretically expected distribution of molecular states. The probability of a state x depends on the total energy $H(q, p)$ of this state. According to Boltzmann, the probability of a state x is proportional to

$$\pi(q, p) \propto \exp(-\frac{1}{k_B T} H(q, p)), \tag{2}$$

where T is the temperature and k_B the Boltzmann constant. In the followings we will always assume that the probability density functions can be normalized on a given position space Ω. In our case, H is separable, i.e. the Boltzmann distribution can be seen as the product distribution $\pi_q \cdot \pi_p$ of a distribution π_q of position coordinates and a distribution π_p of momentum variables. π_q and π_p are strictly positive functions. In this article, the separation of the density function into a position and a momentum part will be understood as follows: Independently from the position state q, the momentum variables are distributed according to π_p for each state $x = (q, p)$ of the system. Equation (2) shows an equilibrium distribution of a molecular ensemble, but it does not provide an equation of motion. Given an initial state x, how will the system evolve in time? Since Newton has formulated the equations of motion, we believe, that given a state x, we can predict the course of the molecular system. This course should somehow resemble (1). MD trajectories,

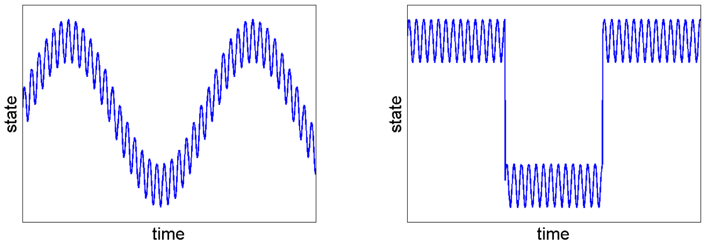

FIGURE 1. *Left.* A dynamical system including different timescales. *Right.* A flip-flop behaviour between two different metastable subsets in state space. The transitions between these sets are rare events. In contrast to the left curve, the transitions can not be described by a continuous dynamics on a different timescale sufficiently.

however, trace isolines of H. Hence, dynamical models have been invented which force the MD trajectories to change between the total energy levels according to the Boltzmann distribution (2). Many researchers have created models for a canonical ensemble dynamics, such that the distribution of simulation data of a single long-time trajectory converges to (2). They have been inspired by the equations of motion (1). Mainly two approaches are used in practise.

1. A deterministic approach: Instead of (1) an alternative similar deterministic dynamical system is defined which converges against Boltzmann distribution. A well-known example is the time-reversible Nosé-Hoover dynamics [1, 2]. Another example is the Berendsen thermostat [3] which does not generate the canonical ensemble exactly. Other time-reversible deterministic thermostats can be found in [4]. It should be mentioned, that the term "deterministic approach" is only of academic interest. From a numerical point of view, the Ljapunov exponent of the dynamical systems is usually very high: Long-time deterministic dynamical systems are chaotic. This is the reason why many researchers prefer molecular dynamics simulations for generating Boltzmann distributed ensembles.
2. A stochastic approach: Beside Smoluchowski [5] and Langevin dynamics [6], the class of hybrid Monte-Carlo methods (HMC) [7] is an example for a stochastic approach towards canonical ensemble dynamics. In HMC, the system is mainly propagated according to (1). Sole exception: After a certain time-span the momentum coordinates are refreshed randomly and a Metropolis-like acceptance step assures the convergence of the system towards (2). Since a total refreshing of momentum variables seems to be unphysical, there are alternative variants of this method. In these variants, momentum variables are more or less conserved, e.g., targeted shadow HMC [8].

The two approaches towards a canonical ensemble dynamics have an important property in common – the Markov property. Given a starting point $x = (q, p) \in \Omega \times \Gamma$, one can determine the probabilities for the possible future evolutions of the system. These propabilities only depend on the starting point x. From this point of view, a time-discretized computation of one of the mentioned dynamical models is nothing else but a realization of a Markov chain in phase space. In the above models, canonical ensemble dynamics try to combine the equation of motion (1) with a correct sampling of states according to (2). Beside possible physical inconsistencies of these models, there is always an unkown additional parameter which defines how fast the trajectories can change between the energy levels of H. From a physical point of view, this parameter determines the quality of the energy transfer of the molecular system with its environment in order to equilibrate temperature. This parameter is difficult to define and often appears arbitrarily. The aim of this article is to avoid analyzing single trajectories in phase space. From our point of view, the choice of a certain thermostated dynamics is not relevant. Instead of applying molecular dynamics, we in the following concentrate on conformation dynamics. The motivation is given by an observation: It is often claimed that molecular dynamical systems include many different timescales, starting from thermal vibration of bonds up to conformational changes of proteins or even larger systems. In our opinion, this statement might be misleading. A conformational change of a molecular system is a rare event, but the transition itself can be very fast, see Figure 1. Thus, on a larger timescale, the conformational changes of the molecular system can better be understood as a flip-flop behaviour [9]. From this flip-flop point of view, the dynamics is described by a matrix of transition probabilities (between "flip" and "flop"). The

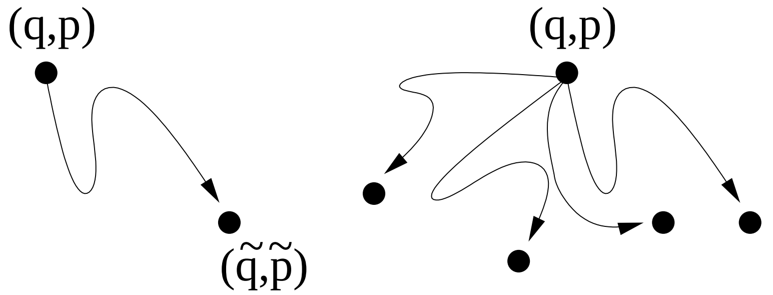

FIGURE 2. *Left.* In the case of (3), Hamiltonian dynamics is deterministic. A given initial state (q,p) leads to a fixed propagated state $(\widetilde{q},\widetilde{p}) = \Phi_{-\tau}(q,p)$. *Right.* In the general case of (4) with a stochastic differential equation or with a Markov chain, the initial state (q,p) is propagated to different states with a different probability. $\Psi_{-\tau}\big(\cdot\,|(q,p)\big)$ is the corresponding probability density function in Ω.

conformations of a molecular system are dynamically metastable subsets[1] of the *position space* [10]. For an analysis of rare events in canonical ensemble dynamics, one has to decompose the position space Ω into dynamically metastable subsets (conformations) and provide a transition probability matrix for the flip-flop behaviour between them[2]. The paper aims at efficiency. It affirms that this analysis can be done without performing any dynamics simulations.

MARKOV PROPERTY AND SCHÜTTE OPERATOR $\mathscr{T}$

In (2), it is not claimed that a single ergodic long-time trajectory samples the phase space according to the Boltzmann distribution[3]. It is only claimed that a given Boltzmann distribution of initial states is conserved by the dynamical system. In this case, (1) may be a valid dynamical model preserving the canonical ensemble distribution. Note that Hamiltonian dynamics (1) is not ergodic in the canonical ensemble.

Conformations as metastable subsets of a molecular system are defined in position space only. Thus, the marginal distribution in Ω with regard to π_q will play an important role. Given a function $f : \Omega \to R$ in position space, the momentum-averaged effect of a dynamics simulation (1) on f is described by Schütte's operator $\mathscr{T}(\tau)$, see [10]:

$$\mathscr{T}(\tau)\, f(q) = \int_\Gamma f(\Pi_q \Phi^{-\tau}(q,p))\, \pi_p(p)\, dp. \tag{3}$$

Equation (3) can be understood as follows: Given an initial state (q,p), a backward Hamiltonian dynamics for a time-interval τ is performed. The new state is denoted as $\Phi^{-\tau}(q,p)$. Via Π_q, this new state is projected to position space. The integral in (3) averages over all possible initial momentum variables with given Boltzmann distribution π_p. In order to write down the operator (3), the Markov property of Hamiltonian dynamics is important. In the Hamiltonian case, the initial state (q,p) exactly determines the final state $\Phi^{-\tau}(q,p)$. Later on we will see that this definition of $\mathscr{T}(\tau)$ leads to an operator which is not time-harmonic. A more general definition of Schütte's operator using the Markov property is given by the *momentum-based tranfer operator*:

$$\mathscr{P}(\tau)\, f(q) = \int_\Gamma \Big(\int_\Omega f(\widetilde{q})\; \Psi_{-\tau}\big(\widetilde{q}|(q,p)\big)\, d\widetilde{q} \Big)\, \pi_p(p)\, dp. \tag{4}$$

In equation (4), the initial state (q,p) determines a probability density function $\Psi_{-\tau}\big(\cdot\,|(q,p)\big)$ for the possible evolutions of the system in position space. For an explanation see Figure 2. $\Psi_{-\tau}$ is a Dirac delta function in the case of a deterministic dynamics. Equation (4) can be used to define a momentum-based transfer operator for any of the dynamical models (deterministic and stochastic) mentioned above, even in the case of a dynamical model

[1] In this article, the meaning of the terms *conformation of a molecular system* and *metastable subset* is different from the common usage in chemistry. Both of the terms denote a subset of the configuration or position space Ω. The term *metastable* means that the subset is *almost stable* with regard to a dynamical process. In chemistry, a metastable subset can be understood as a basin of the potential energy function of the molecular system.

[2] We will see in the following that this is not possible in a strict sense.

[3] Ergodicity is not needed, but from a physical point of view, only an ergodic dynamical model explains self-equilibration of molecular systems.

which is independend from momentum variables – like Smoluchowski dynamics [5]. Going from a trajectory based simulation of molecular systems according to one of the mentioned dynamical models to an analysis of the momentum-based transfer operator is the main step for gaining efficiency. This step leads us from analyzing single trajectories to analyzing a set of Boltzmann distributed trajectories. Instead of choosing a certain dynamical model, we will require certain physical properties for the operator $\mathscr{P}(\tau)$. In order to define $\mathscr{P}$, we already assumed the Markov property for the dynamical model. Furthermore, we will assume a detailed balanced Boltzmann equilibrium distribution and time-harmony for the operator $\mathscr{P}$ in the following. In fact, these properties hold for more than one dynamical model Ψ_τ, see Appendix A.

STATE-BASED VERSUS ENSEMBLE-BASED TRANSITION PROBABILITIES

Hamiltonian dynamics is Markovian in phase space. It is not Markovian in position space. If we apply a projection and go from single states to sets of states in phase space, the Markov property is always lost. The same problem arises when we use the operator $\mathscr{P}$ and a discretization of Ω in order to compute transition probabilities between subsets of Ω [11]. There is always a difference between *state-based transition probabilities* and the *fraction of states which go from set A to set B* in ensemble dynamics. In order to explain this difference, we assume a Markov chain in a finite state space. The transition probabilities of this chain can be expressed by a matrix P. In this example, the vector d is the invariant distribution of P with $d^\top P = d^\top$. Let $\chi_1,\ldots,\chi_k$ denote a set of characteristic vectors defining a decomposition of the state space into subsets. The element (i,j) of the dimension reduction P_c of P onto the given subsets can be written as:

$$P_c(i,j) = d_i^{-1} \chi_i^\top D P \chi_j,$$

where $D = \mathrm{diag}(d)$ is the diagonal matrix of the vector d. The matrix P_c is stochastic: The row sums are 1 and the elements are non-negative. Note that the transition probability from a state q of the set χ_i to the set χ_j depends on q. It is not the same quantity for all $q \in \chi_i$. $P_c(i,j)$ can not be the correct probability, it is an average value. P_c is not the transition matrix of a Markov chain. See also [11] for a detailed example. Whereas, the transition probabilities for single states are different from the entries in P_c, $P_c(i,j)$ provides the correct *fraction of trajectories* going from i to j in one step of the Markov chain, if the initial states are distributed according to d projected to χ_i. This is an ensemble-based point of view. Let us now go back to continuous spaces. Due to the inherent Π_q-projection, $\mathscr{P}$ is an ensemble-based transfer operator. Whenever we speak of *transition probabilities* derived from $\mathscr{P}$ in the following, we think of *the fraction of states which go from A to B*. Conformation dynamics and molecular dynamics are substantially different research areas. In Conformation dynamics we do not aim at a realization of long-time trajectories. Note, however, that we always can formulate a Markov chain in position space on the basis of a given operator $\mathscr{P}(\tau)$. The procedure[4] is as follows: Given a position state $q \in \Omega$, we randomly choose an initial momentum state $p \in \Gamma$ according to π_p. For this state $x = (q,p)$, we compute a realization of the dynamical model $\Psi_{-\tau}\big(\widetilde{q}|(q,p)\big)$ for the given time interval τ and end up with a new position state $\widetilde{q}$. This algorithm is an *interpretation* of $\mathscr{P}$ in terms of a Markov chain $q \longrightarrow \widetilde{q}$. For all mentioned dynamical models, this interpretation in fact provides an ergodic Markov chain with invariant density (2). The physical drawback is that this Markov chain is not interpretable as a trajectory, because it is discontinuous in momentum space after every time step τ. Therefore, it does not provide the “real” dynamics of the system expressed by $\mathscr{P}$ and by its corresponding dynamical model $\Psi_{-\tau}$.

TIME-REVERSIBILITY AND SPECTRAL PROPERTIES OF $\mathscr{P}$

Hamiltonian dynamics (1) is time-reversible. Given an initial state $(q,p) \in \Omega \times \Gamma$ and its propagated state $(\widetilde{q},\widetilde{p}) = \Phi^\tau(q,p)$, the following holds: $(q,-p) = \Phi^\tau(\widetilde{q},-\widetilde{p})$. This fact is used in [10] to show $L^2_{\pi_q}(\Omega)$-self-adjointness of Schütte's operator $\mathscr{T}(\tau)$. An initial state (q,p) and its propagated state $(\widetilde{q},\widetilde{p})$ have the same total energy H and, therefore, the same Boltzmann probability in equation (2), i.e. $\pi_q(\widetilde{q}) \cdot \pi_p(\widetilde{p}) = \pi_q(q) \cdot \pi_p(-p)$. In this equation $\pi_p(\widetilde{p})$ denotes the probability for a transition $\widetilde{q} \longrightarrow q$ and $\pi_p(-p)$ denotes the probability for $q \longrightarrow \widetilde{q}$ (both backward in time).

[4] We often show animations of molecular systems moving according to this kind of Markov chain in our presentations, but we always give a remark that this is a sampling of the position space, and we are only interested in the statistical results of the procedure, not in the special realization of the Markov chain.