High-order averaging schemes with error bounds for thermodynamical properties calculations by MD simulations

Eric Cancès, François Castella, Philippe Chartier, Erwan Faou, Claude Le Bris, Frédéric Legoll, Gabriel Turinici

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

High-order averaging schemes with error bounds for thermodynamical properties calculations by MD simulations

Eric Cancès — François Castella — Philippe Chartier — Erwan Faou — Claude Le Bris — Frédéric Legoll — Gabriel Turinici

N° 4875

Juillet 2003

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4875--FR+ENG

INRIA
ROCQUENCOURT

High-order averaging schemes with error bounds for thermodynamical properties calculations by MD simulations

Eric Cancès* , François Castella† , Philippe Chartier‡ , Erwan Faou‡ , Claude Le Bris* , Frédéric Legoll*§ , Gabriel Turinici*

Thème 4 — Simulation et optimisation
de systèmes complexes
Projets MICMAC et ALADIN

Rapport de recherche n° 4875 — Juillet 2003 — 21 pages

Abstract: We introduce high-order averaging formulae for the computation of statistical averages on the basis of the numerical simulation of long-time molecular dynamics trajectories. Using these formulae, we improve the convergence rate of the time average with respect to the computational effort. We provide some numerical examples that show the efficiency of our scheme. When trajectories are approximated using symplectic integration schemes (such as velocity Verlet), we give error bounds that provide guidelines to choose the parameters of the computation in order to reach a given desired accuracy in the most efficient manner.

Key-words: Molecular dynamics, statistical average computation, high-order averaging schemes, error bounds, ergodic theorem

* CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, Cité Descartes, 77455 Marne-la-Vallée Cedex 2, France, and MICMAC, INRIA Rocquencourt
† IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France
‡ ALADIN, INRIA Rennes
§ EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart, France

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 —Télécopie : +33 1 39 63 53 30

Schémas d'ordre élevé pour le calcul de moyennes en dynamique moléculaire

Résumé : Nous introduisons des formules d'ordre élevé pour le calcul de moyennes statistiques à partir de simulations numériques de trajectoires de dynamique moléculaire en temps long. Ces formules permettent d'accélérer la convergence des moyennes temporelles vers les moyennes d'ensembles thermodynamiques. Cette approche est testée sur plusieurs exemples numériques qui montrent son efficacité. Lorsque les trajectoires sont calculées par un algorithme symplectique (comme velocity Verlet), nous donnons des estimateurs d'erreur qui guident le choix des paramètres de la simulation, afin d'obtenir des résultats avec une précision donnée de la manière la plus efficace possible.

Mots-clés : Dynamique moléculaire, calcul de moyennes statistiques, schémas d'ordre élevé pour le calcul de moyennes, estimation d'erreur, théorème ergodique

1 Introduction

The properties of a given physical system at thermodynamical equilibrium (radial distribution functions, free energies, transport coefficients, ...) can be computed as averages of some observables over the phase space of one representative underlying microscopic system [1, 2].

In most applications of interest, this microscopic system is composed of a high number of particles (currently studied biological systems involve more than 100,000 atoms), so that the dimension of the phase space (which is $6M$, where M is the number of particles) is also very large, making the computation of the averages a challenging task.

Two families of methods are currently used in order to sample the phase space according to the convenient probability density (which depends itself of the thermodynamic ensemble in which the calculation is performed: NVE, NVT, NPT, ...): Monte Carlo methods [1, 2] on the one hand, and Molecular Dynamics (MD) on the other hand. In the latter case, the time evolution of the microscopic system is simulated (possibly coupled with a bath, such as in NVT, NPT, ...calculations); under the ergodicity assumption, the trajectory of the system properly samples the phase space and the time average of the observable over the trajectory converges toward the ensemble average of the observable when the simulation time goes to infinity [3].

When time averages are estimated by MD calculations, a numerical scheme is needed to compute the evolution of the system in the phase space. The resulting numerical trajectory deviates dramatically from the exact one in the long time limit. Furthermore, the simulated trajectory is of course of finite length. In practice, we thus are far from the ideal object of an exact trajectory of infinite length. It is however of common belief and commonly observed that, in many situations, satisfying results are obtained for the time averages, despite the not very high trajectory accuracy and the not tremendously long simulation time.

In this article, we first provide an apparently new method for calculating the time average. It allows one either to significantly speed up the convergence for a given accuracy or to achieve a better accuracy at a given computational cost. In addition, we contribute to the above common belief through an *a priori* error estimation which provides an upper bound of the error between the actual ensemble average and the numerically computed time average, function of the numerical parameters of the calculation, i.e. the (finite) integration time step and the (finite) trajectory length. This estimation also involves the order of the numerical scheme used to compute the trajectory as well as some parameters related to the time averaging method. We show in particular that the error due to the finiteness of the simulation time T decays as $1/T$ if the time average is computed by discretizing the standard formula $\frac{1}{T}\int_0^T A(t)\,dt$ ($A(t)$ denotes the value of the observable at time t) whereas this component of the error decreases as $1/T^k$ where k is a chosen positive integer if the time average is evaluated with the formula $\int_0^T A(t)\,f_{k,T}(t)\,dt$ where $f_{k,T}(t)$ is some filtering function whose expression is provided in Sec. 2.

We only deal here with the NVE thermodynamical ensemble. The equations of motion which correspond to this ensemble are the Newton equations (the "physical" dynamics of the system), which can be recast as a Hamiltonian dynamical system. However, similar results are likely to hold within other ensembles, provided that the underlying dynamical system is Hamiltonian; that is the case when Nosé-Poincaré thermostats [4] are used to perform computations in the NVT ensemble. We are currently working on such extension of our approach.

Under some very strong hypotheses collected in Appendix 5, including in particular that the Hamiltonian system is integrable, we are able to rigorously establish the error estimate mentioned above and set in Sec. 2. The rather intricate and mathematically demanding proof of this estimate can be read in a companion article [5]. We provide in Secs. 3.1 and 3.2 numerical examples on integrable systems that show that our estimate is indeed sharp. Of course, very few of the systems used in chemistry or physics are integrable. Fortunately, even if we are not able to prove it rigorously, our estimate also seems to hold true in more realistic situations, as it is shown in Sec. 3.3, where convincing results on a system of 60 particles in 3D interacting through a truncated Lennard-Jones potential are provided. Lastly, we investigate in Sec. 3.4 the example of a 2D particle in a double well potentiel, which is a toy model mimicking a larger system whose potential energy surface presents several basins corresponding to metastable states. Transitions between these states are possible but are rare events in comparison with the faster characteristic time scale of the system. It is shown on this simple example that if the energy of the particle is just above the saddle point energy of the potential, the rate of convergence of the time averages calculated by MD is slowed down to $1/\sqrt{T}$, due to the quasi-stochastic transition process between the two basins of the double well potential. It is also shown that, if the energy of the particle is smaller or much larger than the saddle point energy, the time averages converge, but not toward the ensemble average. However, if one uses several initial conditions and compute a time average for each of them, it is shown that the mean value of these time averages is a good approximation of the ensemble average.

2 Main setting and result

Let us consider M particles in 3D. Each particle i is described by its position $\mathbf{q}_i \in \mathbf{R}^3$, its momentum $\mathbf{p}_i \in \mathbf{R}^3$ and its mass m_i. Let $H(\mathbf{q},\mathbf{p})$ be the Hamiltonian of the system, defined on $\mathbf{R}^{3M} \times \mathbf{R}^{3M}$ by

$$H(\mathbf{q},\mathbf{p}) = \sum_{i=1}^{M} \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{q}_1,\ldots,\mathbf{q}_M),$$

where V is the interaction potential, and $\mathbf{q}$ and $\mathbf{p}$ are notations for $(\mathbf{q}_1, \ldots, \mathbf{q}_M)$ and for $(\mathbf{p}_1, \ldots, \mathbf{p}_M)$ respectively. The Hamiltonian dynamical system associated to H is given by

$$\left\{ \begin{array}{rcl} \dfrac{d\mathbf{q}}{dt} & = & \dfrac{\partial H}{\partial \mathbf{p}}(\mathbf{q}(t), \mathbf{p}(t)), \\ \dfrac{d\mathbf{p}}{dt} & = & -\dfrac{\partial H}{\partial \mathbf{q}}(\mathbf{q}(t), \mathbf{p}(t)). \end{array} \right. \tag{1}$$

We suppose that the dynamical system (1) has $3M$ invariant functions $I_j(\mathbf{q}, \mathbf{p})$, $j = 1, \ldots, 3M$. Let us recall that $I_j(\mathbf{q}, \mathbf{p})$ is an invariant function if $I_j(\mathbf{q}(t), \mathbf{p}(t))$, where $(\mathbf{q}(t), \mathbf{p}(t))$ is solution of (1), is constant with respect to time. Of course, $H(\mathbf{q}, \mathbf{p})$ is one of these invariants, since the energy is preserved by the dynamics. We denote by

$$S(\mathbf{q}, \mathbf{p}) = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbf{R}^{3M} \times \mathbf{R}^{3M} \text{ such that } \forall j \in [1, 3M],\ I_j(\mathbf{x}, \mathbf{y}) = I_j(\mathbf{q}, \mathbf{p}) \right\} \tag{2}$$

the level set of the invariant functions $\{I_j\}_{1 \leq j \leq 3M}$ containing the phase space point $(\mathbf{q}, \mathbf{p})$.

Let $(\mathbf{q}_0, \mathbf{p}_0)$ be an initial condition, and let $A(\mathbf{q}, \mathbf{p})$ be an observable whose average is well-defined in the NVE ensemble. As we suppose that the dynamical system has $3M$ invariants, the actual trajectory starting from the initial point $(\mathbf{q}_0, \mathbf{p}_0)$ remains for all times on the level set $S(\mathbf{q}_0, \mathbf{p}_0)$. Thus, the relevant ensemble average of the observable A reads

$$\langle A \rangle = \frac{\displaystyle\int_{S(\mathbf{q}_0, \mathbf{p}_0)} A(\mathbf{q}, \mathbf{p})\, d\mu(\mathbf{q}, \mathbf{p})}{\displaystyle\int_{S(\mathbf{q}_0, \mathbf{p}_0)} d\mu(\mathbf{q}, \mathbf{p})} \tag{3}$$

where $d\mu(\mathbf{q}, \mathbf{p})$ is the invariant measure on $S(\mathbf{q}_0, \mathbf{p}_0)$.

Under some hypotheses collected in Appendix 5, one can prove [5, 6] the so-called ergodic theorem, that is

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T A(\mathbf{q}(t), \mathbf{p}(t))\, dt = \langle A \rangle, \tag{4}$$

where $(\mathbf{q}(t), \mathbf{p}(t))$ is the trajectory of the system (1) starting from the initial point $(\mathbf{q}_0, \mathbf{p}_0)$. As an approximation to the average (3), it is a standard approach to use the discrete sum

$$\langle A \rangle_{num}^{Rie}(\delta t, T) = \frac{1}{\mathcal{N}} \sum_{j=0}^{\mathcal{N}-1} A\left(\mathbf{q}_j, \mathbf{p}_j\right), \tag{5}$$

where $(\mathbf{q}_j, \mathbf{p}_j)_{j=0}^{\mathcal{N}-1}$ is the numerical trajectory given by an integration scheme of time step δt applied on Eqs. (1), and where $T = \mathcal{N} \delta t$. However, as the convergence in (4) occurs [6] only at speed $1/T$, we cannot expect a faster convergence rate of (5) toward $\langle A \rangle$.

In order to improve this asymptotic convergence rate, we suggest to replace the uniformly-weighted time average $\frac{1}{T}\int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\,dt$ by a filtered time average of the form

$$\int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\,f_T(t)\,dt, \tag{6}$$

where f_T acts as a filtering function of the signal $t \mapsto A(\mathbf{q}(t),\mathbf{p}(t))$. For instance, denoting by $I_1(t_1) = \frac{2}{T}\int_{t_1}^{t_1+T/2} A(\mathbf{q}(t),\mathbf{p}(t))\,dt$ the time average in the interval $[t_1, t_1+T/2]$ (with $t_1 \in [0,T/2]$), one can consider the average of $I_1(t_1)$ over t_1, i.e. $\langle A\rangle^{(2)}(T) = \frac{2}{T}\int_0^{T/2} I_1(t_1)dt_1$. Let us define the function $\alpha_{[0,\mathcal{T}]}$ by

$$\alpha_{[0,\mathcal{T}]} \ : \ t \in \mathbf{R} \quad \mapsto \quad 1/\mathcal{T} \ \text{ if } t \in [0,\mathcal{T}], \quad 0 \text{ otherwise.}$$

Then $\langle A\rangle^{(2)}(T)$ can be recast into (6), with $f_T \equiv f_{2,T}$, where $f_{2,T}$ is the convolution of $\alpha_{[0,T/2]}$ with itself. More generally, denoting by $f_{k,T}$ the k-th convolution of the function $\alpha_{[0,T/k]}$ with itself, we define the time average $\langle A\rangle^{(k)}(T) = \int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\,f_{k,T}(t)\,dt$. We can prove [5], under the same assumptions as above, that

$$\langle A\rangle^{(k)}(T) = \langle A\rangle + O\left(\frac{1}{T^k}\right). \tag{7}$$

Thus, for any **arbitrary** positive integer k, it is possible to design an averaging scheme which converges to the ensemble average at speed $1/T^k$.

The discrete version of $\langle A\rangle^{(k)}(T)$ is

$$\langle A\rangle^{(k)}_{num}(\delta t, T) = \frac{1}{N^k}\sum_{j=0}^{k(N-1)} C(k,N-1,j)A\left(\mathbf{q}_j,\mathbf{p}_j\right), \tag{8}$$

where $(\mathbf{q}_j,\mathbf{p}_j)_{j=0}^{kN}$ is the numerical trajectory given by an integration scheme of time step δt, $T = kN\,\delta t$ and where $C(k,N-1,j)$ is the number of k-tuples of positive integers $j_1,\ldots,j_k$ whose sum is fixed to be j and which all belong to $[0,N-1]$. These numbers can be computed by recursion, in such a way that their computation is almost free in comparison to the computation of the trajectory $(\mathbf{q}_j,\mathbf{p}_j)_{j=0}^{kN}$.

The formula (8) is a generalization of (5) since $\langle A\rangle^{(k=1)}_{num}(\delta t,T) = \langle A\rangle^{Rie}_{num}(\delta t,T)$. To the best of our knowledge, the scheme (8) does not seem to be used in Molecular Dynamics simulations, or at least to be reported on in the literature. Of course, any information or suggestion from the reader, particularly on this point or on more general issues, will receive our best attention.

In addition, if a symplectic scheme of order r_0 is used to compute the trajectory (recall that, for instance, the velocity Verlet scheme is symplectic and of order $r_0 = 2$), we can prove [5] the following error estimate:

$$| \langle A \rangle_{num}^{(k)}(\delta t, T) - \langle A \rangle | \leq C \left(\frac{1}{T^k} + \delta t^{r_0} \right) \tag{9}$$

for some constant C which does not depend on T or δt.

We underline that the polynomial decay of the error with respect to the total time of simulation T is made possible by the fact that the values $A\left(\mathbf{q}_j, \mathbf{p}_j\right)$ are correlated in time, in the sense that they are obtained by the simulation of one single trajectory. When one uses a Monte Carlo method, the values $A\left(\mathbf{q}_j, \mathbf{p}_j\right)$ are not correlated any more, and the error only decays at the universal rate $1/\sqrt{T}$ (see also the example of the double well in Sec. 3.4).

The proof of (9), together with details on the filter used in formula (8), may be read in a companion article [5]. Current efforts of research are directed towards designing other filters that might further improve the efficiency of our approach.

3 Numerical examples

We now turn to numerical examples that show:

- the sharpness of the estimate (9) for integrable Hamiltonian systems satisfying (at least some of) the assumptions under which the estimate can be rigourously proved.
- that the estimate (9) also holds for non-integrable systems of practical interest.

We first present results in the case of a collection of independent harmonic oscillators (see Sec. 3.1), then on the Kepler problem (see Sec. 3.2). Both systems are integrable, and are likely to fall within the hypotheses of our theoretical result. In Sec. 3.3, we study a system of 60 particles subjected to the Lennard-Jones potential. Finally, in Sec. 3.4, we study a particle submitted to a double well potential in 2D. This toy-model is representative of many systems studied in chemistry and physics, in the sense that the potential has several basins in which the system can stay for a long time. The exploration of the whole phase space needed to accurately compute thermodynamic properties is in these cases particularly difficult.

For the sake of completeness, and to show that our approach is insensitive to the numerical symplectic integration scheme, we have used three different algorithms [7] of different orders to integrate the dynamical system (1):

- the standard velocity Verlet algorithm, which is, as recalled above, symplectic and of order 2; we denote by $\phi_{\delta t}^{VV}$ a step of the algorithm of time step δt: $(\mathbf{q}_{n+1}, \mathbf{p}_{n+1}) = \phi_{\delta t}^{VV}(\mathbf{q}_n, \mathbf{p}_n)$;

- compositions of the velocity Verlet algorithm with itself so that the order of the algorithm increases to 4; namely,

$$(\mathbf{q}_{n+1}, \mathbf{p}_{n+1}) = \phi_{\delta t}^{(4)}(\mathbf{q}_n, \mathbf{p}_n) = \phi_{b_1\delta t}^{VV} \circ \phi_{b_0\delta t}^{VV} \circ \phi_{b_1\delta t}^{VV}(\mathbf{q}_n, \mathbf{p}_n)$$

  with $b_1 = 1/(2 - 2^{1/3})$ and $b_0 = 1 - 2b_1$.

- compositions of the velocity Verlet algorithm with itself so that the order of the algorithm increases to 6; namely,

$$(\mathbf{q}_{n+1}, \mathbf{p}_{n+1}) = \phi_{\delta t}^{(6)}(\mathbf{q}_n, \mathbf{p}_n) = \phi_{c_3\delta t}^{VV} \circ \phi_{c_2\delta t}^{VV} \circ \phi_{c_1\delta t}^{VV} \circ \phi_{c_0\delta t}^{VV} \circ \phi_{c_1\delta t}^{VV}\phi_{c_2\delta t}^{VV} \circ \phi_{c_3\delta t}^{VV}(\mathbf{q}_n, \mathbf{p}_n)$$

  with $c_1 = -1.17767998417887$, $c_2 = 0.235573213359357$, $c_3 = 0.784513610477560$ and $c_0 = 1 - 2(c_1 + c_2 + c_3)$.

In all the figures below, we plot either the numerical time average $\langle A \rangle_{num}^{(k)}(\delta t, T)$ of an observable A or the error $\mid \langle A \rangle_{num}^{(k)}(\delta t, T) - \langle A \rangle \mid$ as a function of T. Consequently, the computational cost is the same for all points vertically aligned.

3.1 Collection of harmonic oscillators

In this section, we consider a system of five particles in 1D whose dynamics is described by the Hamiltonian

$$H = \sum_{i=1}^{5} \frac{p_i^2}{2} + \omega_i^2 \frac{q_i^2}{2},$$

with the following pulsations:

$$\omega_1 = 1.0, \quad \omega_2 = 8.01256, \quad \omega_3 = 12.25245, \quad \omega_4 = 17.234, \quad \omega_5 = 20.98765.$$

Let us for instance study the average of $A = p_4^2$. The exact value of the ensemble average can be obtained by analytical calculation: $\langle A \rangle = 0.5$.

Let us briefly check whether the hypotheses needed to prove our estimates are satisfied. The integration algorithm is symplectic. The analyticity and integrability assumptions on the Hamiltonian function (see Appendix 5.1) are satisfied. We are able neither to prove nor to disprove the diophantine assumption (see Sec. 5.2).

We use the formula (8) with different values of k ($k = 1$ to 4), for a given δt ($\delta t = 0.025$), and with the order-6 algorithm. Clearly (see Fig. 1), increasing k gives a better convergence.

3.2 The Kepler problem

We next turn to the simulation of the Kepler problem. We consider the Hamiltonian

$$H = \frac{\mathbf{p}^2}{2} - \frac{1}{\mid \mathbf{q} \mid},$$

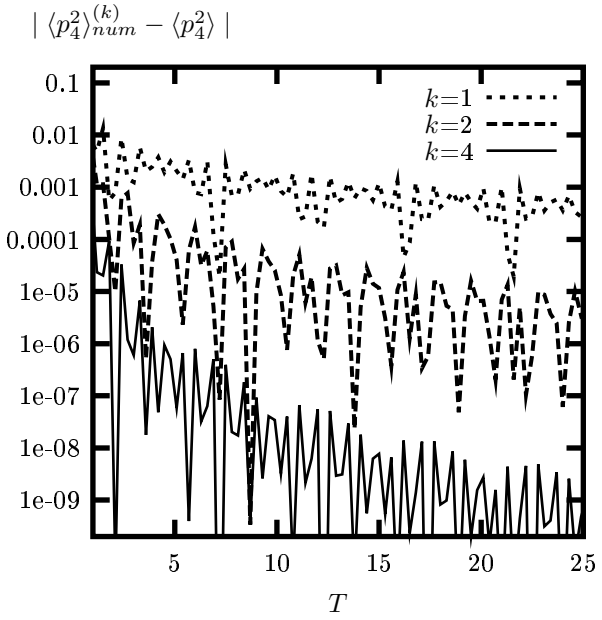

Figure 1: Convergence of $\langle p_4^2 \rangle_{num}^{(k)}(\delta t, T)$ to $\langle p_4^2 \rangle$ for different values of k, in the case of a collection of 5 harmonic oscillators. Initial conditions are $q_1 = 4$, $p_1 = -1.2$, $q_i = 0$ and $p_i = 1$ for $i = 2, \ldots, 5$.

describing a 3D particle in an attractive Coulomb potential.

Let us for instance study the average of $A = \mid \mathbf{q} \mid$. As in the previous example, the ensemble average can be computed analytically: $\langle \mid \mathbf{q} \mid \rangle = \left(3 + 2\, E_0 L_0^2\right) / (4 \mid E_0 \mid)$, where E_0 is the initial (negative) energy and L_0 is the norm of the initial kinetic momentum. The simulation time step is $\delta t = 0.01$, and we use the order-6 algorithm. As for the harmonic oscillators, the Hamiltonian function is analytical and integrable, the integration scheme is symplectic, but we are able neither to prove nor to disprove the diophantine assumption (see Sec. 5.2).

We compute the averages with k ranging from 1 to 5. We see that, as k increases, the amplitude of the oscillations at the beginning of the simulation increases (see Fig. 2). This fact is generic, and not restricted to the Kepler problem. However, our approach focuses on long simulation times. We see here that, for $T \geq 40$ (approximately beyond 4 revolution periods), the larger k, the faster the convergence (see Fig. 3).

On this case, we are able to check that the estimate (9) is sharp. For this purpose, we study the dependence of the error with respect to T and δt in order to check whether there exist constants $C_1(k)$ and $C_2(r_0)$ such that

$$\mid \langle A \rangle_{num}^{(k)}(\delta t, T) - \langle A \rangle \mid \approx \frac{C_1(k)}{T^k} + C_2(r_0) \delta t^{r_0}.$$

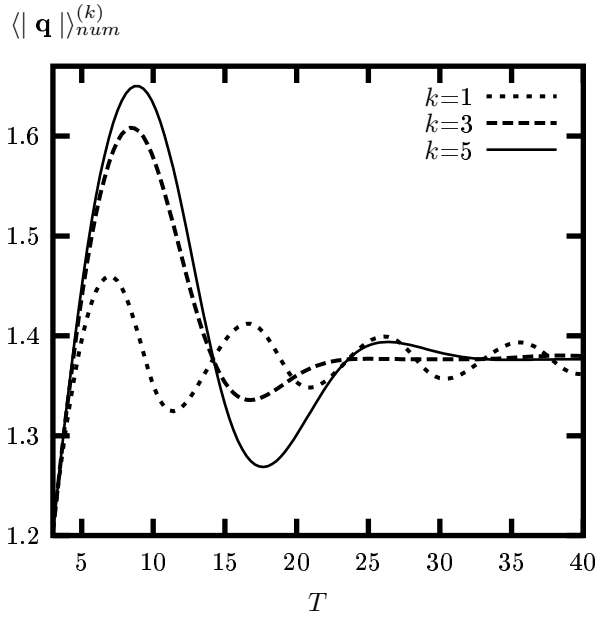

Figure 2: Oscillations of $\langle | \mathbf{q} | \rangle_{num}^{(k)}(\delta t, T)$ at the beginning of the simulation (Kepler problem), for different values of k. Initial conditions are $\mathbf{q} = (0.9, 0, 0)$ and $\mathbf{p} = (0, 1.1, 0.5)$.

We first try to determine $C_1(k)$. To do this, we work with very small values of δt and not too large values of T, such that the leading term of the error is $C_1(k)/T^k$. We compute the rescaled error $T^k \mid \langle A \rangle_{num}^{(k)}(\delta t, T) - \langle A \rangle \mid$. If our estimate is sharp, this quantity should be bounded independently of T. We find indeed that for $k = 1$, the rescaled error is bounded by 0.6, **independently of** r_0 **and** δt.

Proceeding in the same way for the other values of k, we obtain

$$C_1(k=1) \approx 0.6, \quad C_1(k=2) \approx 14.4, \quad C_1(k=3) \approx 212,$$
$$C_1(k=4) \approx 8270, \quad C_1(k=5) \approx 250,000.$$

Not surprisingly, the values of $C_1(k)$ quicky increase with k (this is in agreement with the fact that the oscillations at the beginning of the simulation are larger for larger values of k).

In order to determine the values of $C_2(r_0)$, we proceed conversely and compute a long-time trajectory with a not too small value of δt. For $k > 1$, as T is large, the term $C_1(k)/T^k$ is very small with respect to the term $C_2(r_0)\delta t^{r_0}$, so the error is approximately given by $C_2(r_0)\delta t^{r_0}$. We find values for $C_2(r_0)$ which depend **neither on** k **nor on** T and are given by

$$C_2(r_0 = 6) = 0.163, \quad C_2(r_0 = 4) = 1.09, \quad C_2(r_0 = 2) = 0.622.$$

These numerical experiments therefore confirm the sharpness of the *a priori* estimate (9).

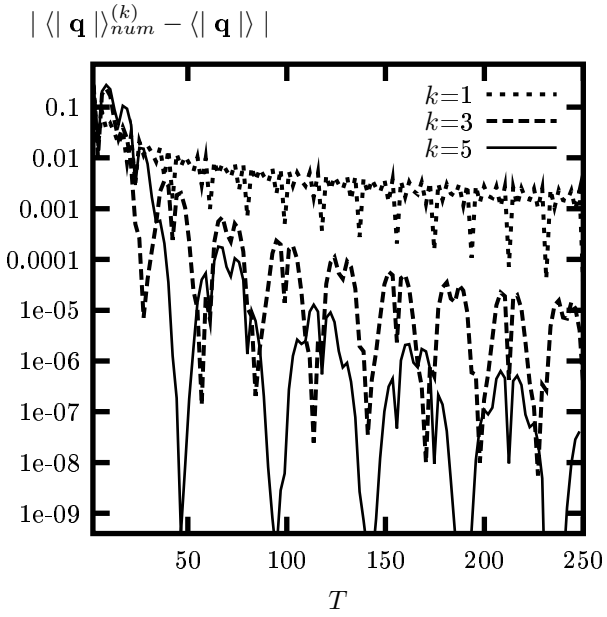

Figure 3: Convergence of $\langle|\mathbf{q}|\rangle_{num}^{(k)}(\delta t, T)$ to $\langle|\mathbf{q}|\rangle$ for different values of k, in the case of the Kepler problem. Initial conditions are $\mathbf{q} = (0.9, 0, 0)$ and $\mathbf{p} = (0, 1.1, 0.5)$.

3.3 60 particles in a truncated Lennard-Jones potential

In this part, we study a system of $M = 60$ particles in 3D interacting through a truncated Lennard-Jones potential. The Hamiltonian of the system reads

$$H = \sum_{i=1}^{M} \frac{\mathbf{p}_i^2}{2} + \sum_{i=1}^{M} \sum_{j>i} V(|\mathbf{q}_j - \mathbf{q}_i|)$$

with $V_{LJ}(z) = \frac{4}{z^{12}} - \frac{4}{z^6}$ and $V(z) = V_{LJ}(z) - V_{LJ}(z_c) - V'_{LJ}(z_c)(z - z_c)$ if $z \leq z_c$, 0 otherwise. Thus, both the potential and the forces are continuous at the cutoff radius $z_c = 3.08$. We apply periodic boundary conditions (the simulation box is $[-3.2; 3.2]^3$) and use the minimum image convention [1]. Initial conditions are such that the density of the system is 0.229 and the average of the kinetic temperature is 2.50. With such an initial condition, the system remains in a liquid phase. Hypotheses collected in Appendix 5 are not satisfied (in particular, there are not enough invariant functions). The observable under examination is the mean distance between particles,

$$A(\mathbf{q}_1, \ldots, \mathbf{q}_M) = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j>i} |\mathbf{q}_j - \mathbf{q}_i| .$$

We work with the velocity Verlet algorithm, using a time step of $\delta t = 5.10^{-3}$.

Once again, we see that, for large enough simulation times T, the larger k, the faster the convergence (see Fig. 4).

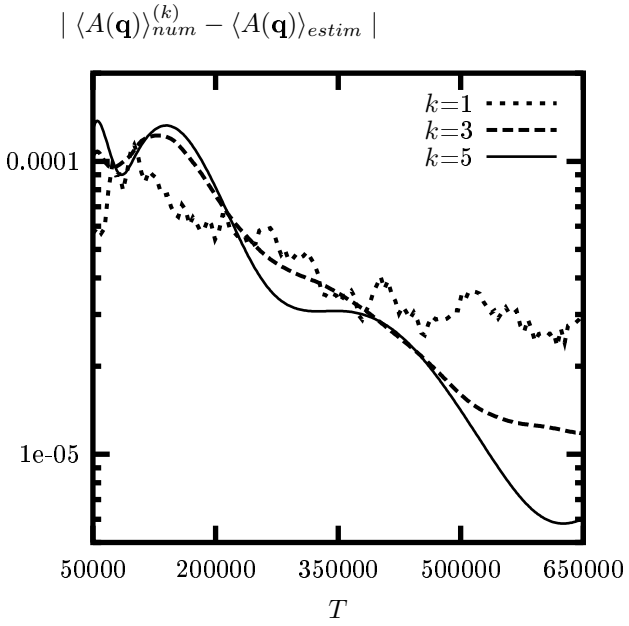

Figure 4: Convergence of $\langle A(\mathbf{q}) \rangle_{num}^{(k)}(\delta t, T)$ to $\langle A(\mathbf{q}) \rangle_{estim}$ for different values of k, in the case of the Lennard-Jones system (the observable is the mean distance between particles).

3.4 One particle in a double well potential

In this part, we deal with the Hamiltonian

$$H = \frac{p_x^2}{2} + \frac{p_y^2}{2} + V(q_x, q_y),$$

where

$$V(q_x, q_y) = (q_x^2 - 1)^2 + (q_y + q_x^2 - 1)^2,$$

describing a 2D particle in a double well potential. The potential V goes to $+\infty$ at infinity and has three critical points: two global minima located at $(\pm 1, 0)$, at which $V(\pm 1, 0) = 0$, and one saddle point located at $(0, 1)$, at which $V(0, 1) = 1$.

The behaviour of time averages strongly depends on the value of the energy E_0 of the particle with respect to the saddle point energy (here equal to 1). Three different regimes can be identified:

1. for E_0 smaller than 1, the trajectory only explores a single basin. The time averages converge to a limit and using the filtering function $f_{k,T}$ improves the convergence rate from $1/T$ to $1/T^k$ (for an arbitrary k) as in the previous examples;

2. for E_0 much larger than 1 (say $E_0 \geq 5$), the particle is so energetic that it does not really "feel" the barrier and the convergence of the time averages is similar to the one observed in the first case;

3. for E_0 a little larger than 1, the system really "feels" the presence of two distinct basins and one does not observe any convergence of the time averages for simulation times of the same order of magnitude (or even 10,000 times longer) as those which lead to convergence in the previous two cases (see Fig. 5).

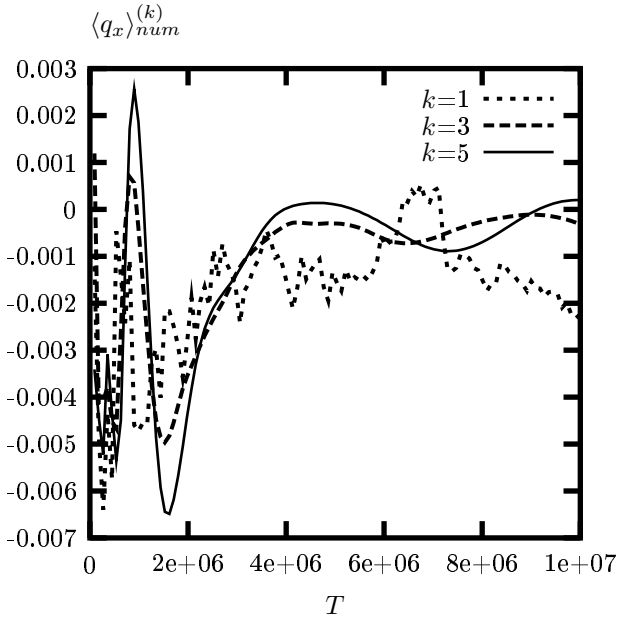

Figure 5: Evolution of $\langle q_x \rangle_{num}^{(k)}(\delta t, T)$ as a function of T (double well potential, energy of 1.25 close to the barrier energy), for a time step $\delta t = 0.05$ and 2.10^8 time steps.

Actually, as one can see from Fig. 6, the particle spends "a long time" in one basin, then quickly undergoes a transition into the other basin, in which it spends another "long time", and so on. In this simple example, the locations of the two basins are known and the mean exit time can be easily estimated from MD simulations. These data can be used to parametrize a two-state Markov chain model [8] mimicking the transition between the two basins. A comparison between the convergence rate of the time averages (8) computed by MD on the one hand and by the so-obtained Markov chain model on the other hand, for the

observable q_x (the average of which is zero due to the symmetry of the potential), is reported on Fig. 7. The good quantitative agreement between the two convergence rates confirms that the bottleneck in the computation of averages by brute force MD simulations is the presence of several well-separated basins; in this case, the convergence rate is no longer in $1/T$ or better, but falls down to $1/\sqrt{T}$, at least in the range of today accessible simulation times, due to the quasi-stochastic transitions between the basins.

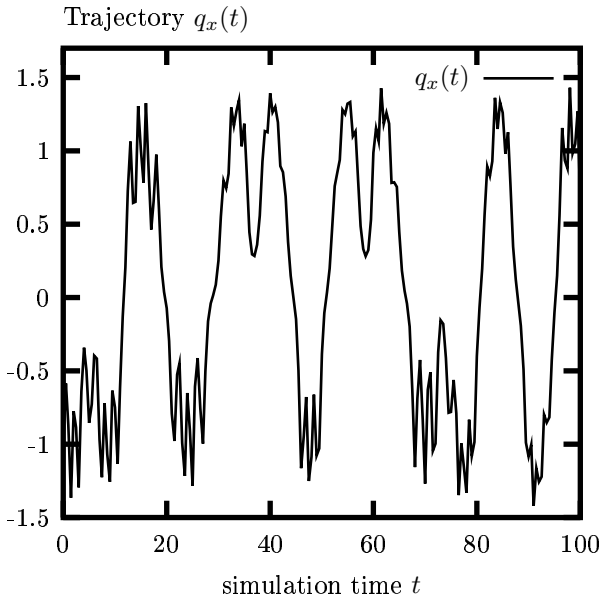

Figure 6: Evolution of $q_x(t)$ as a function of the time t (double well potential, energy of 1.25 close to the barrier energy, time step $\delta t = 0.01$).

Let us end this section by a discussion on the ergodicity of the 2D double well system under study. It is easy to check that the isoenergetic surface $H(q_x, q_y, p_x, p_y) = E_0$ can be parametrized by three angles:

$$\begin{aligned} q_x &= \pm\sqrt{1+\sqrt{E_0}\cos(\theta)}, & q_y &= \sqrt{E_0}\,(\sin(\theta)\cos(\phi)-\cos(\theta)), \\ p_x &= \sqrt{2E_0}\,\sin(\theta)\sin(\phi)\cos(\psi), & p_y &= \sqrt{2E_0}\,\sin(\theta)\sin(\phi)\sin(\psi), \end{aligned} \tag{10}$$

with $(\theta, \phi, \psi) \in [0, \theta_M] \times [0, \pi] \times [0, 2\pi]$, where $\theta_M = \pi$ if $E_0 \leq 1$ and $\theta_M = \arccos(-1/\sqrt{E_0})$ if $E_0 > 1$. It is therefore possible to compute numerically the NVE ensemble average of any (regular) observable with a high accuracy. It reads

$$\langle A \rangle(E_0) = \int_{[0,\theta_M]\times[0,\pi]\times[0,2\pi]} A\,(\theta,\phi,\psi)\; w(\theta,\phi)\; d\theta\, d\phi\, d\psi, \tag{11}$$

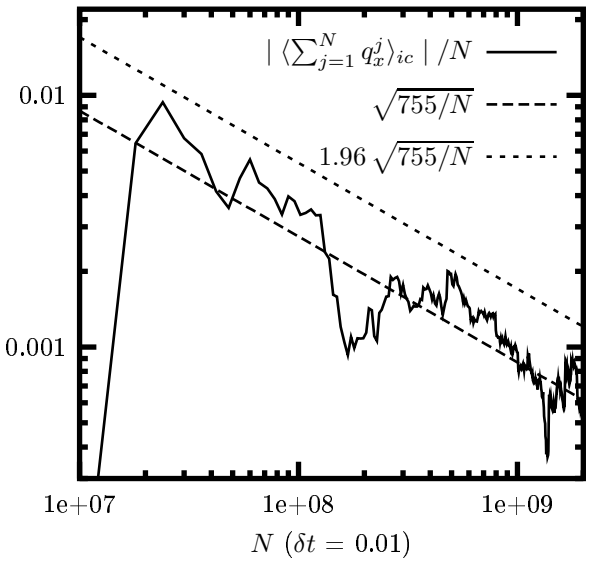

Figure 7: Estimation of the convergence of $\frac{1}{N}\sum_{j=1}^{N} q_x^j$ to 0 with respect to N with a Markov chain model (double well potential, energy of 1.25 close to the barrier energy, $\delta t = 0.01$; $\langle\cdot\rangle_{ic}$ denotes average over 16 initial conditions).

where $w(\theta,\phi) = C\,\dfrac{\sin(\phi)\sin(\theta)^2}{\sqrt{1+\sqrt{E_0}\cos(\theta)}}$ is the invariant probability density of the dynamical system on (θ,ϕ,ψ) (C is the normalizing constant). In Table 1, we compare this ensemble average with several time averages computed from trajectories with various initial conditions. In general, these time averages are different from the ensemble average. This difference can be explained either by the fact that the isoenergetic surface has invariant subsets (this may be due to the existence of a second invariant that we did not manage to identify), or by the fact that the system is ergodic but on very long time scales, much longer than the simulation times we can afford.

From a practical point of view, one way to solve this issue is to compute the mean value of the time averages over several different initial conditions. First, as in a Monte Carlo method, we choose many initial conditions $(\mathbf{q}_i,\mathbf{p}_i)_i$ on the surface of constant energy E_0, according to the probability density $w(\theta,\phi)$ (to do so, we used the Acceptance-Rejection method [9]). For all these initial conditions, we compute a MD trajectory and a time average, and then the mean value $\mu(E_0)$ and the standard deviation $\sigma_{MC+MD}(E_0)$ of these data. Results are reported in Fig. 8: this method yields a good approximation of the ensemble average.

Observable A	$\langle A\rangle(E_0 = 0.5)$	$\langle A\rangle_{num}^{(k=5)}(\mathbf{q}_1, \mathbf{p}_1)$	$\langle A\rangle_{num}^{(k=5)}(\mathbf{q}_2, \mathbf{p}_2)$	$\langle A\rangle_{num}^{(k=5)}(\mathbf{q}_3, \mathbf{p}_3)$
q_x	-0.94459	-0.93118	-0.95585	-0.92386
q_x^2	0.92843	0.90077	0.95170	0.88565
q_x^4	0.98964	0.92855	1.04074	0.89565
q_y	0.071562	0.099221	0.048292	0.114346
q_y^2	0.25517	0.14298	0.34725	0.085192
p_x^2	0.24482	0.33949	0.16794	0.38717
p_x^2	0.24482	0.14615	0.32574	0.095278
$V(q_x, q_y)$	0.25517	0.25717	0.25315	0.25878

Table 1: Computed values for the ensemble average and the time average of different observables (double well potential). The ensemble averages are computed from (11). For the time averages, computed from (8), initial conditions are $(\mathbf{q}_1, \mathbf{p}_1) = (-1, 0.5; 0.5, -0.5)$, $(\mathbf{q}_2, \mathbf{p}_2) = (-0.9, 0.772\ldots; 0, -0.5)$ and $(\mathbf{q}_3, \mathbf{p}_3) = (-1.05, 0.549; 0.3606\ldots, 0)$, corresponding to the energy 0.5, the time step is $\delta t = 0.005$. We checked that a simulation of length $T = 5000$ was long enough. Time average values do not depend on k.

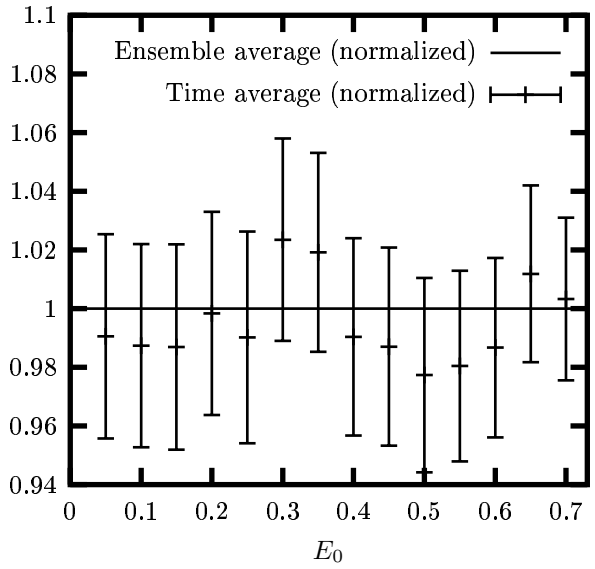

Figure 8: Comparison between the ensemble average of $A = q_y^2$, computed from (11), and the expected time average (error bars correspond to a 95% confidence interval), on the double well potential, for different energy levels E_0. All quantities have been normalized by the ensemble average $\langle q_y^2\rangle(E_0)$.

The mean value $\mu(E_0)$ converges to the ensemble average $\langle A \rangle$ at rate $\sigma_{MC+MD}(E_0)/\sqrt{M}$, where M is the number of initial conditions. If a standard Monte Carlo method is used, the rate of convergence is $\sigma_{MC}(E_0)/\sqrt{M}$, where $\sigma_{MC}(E_0)$ is the standard deviation of A with respect to the probability density $w(\theta, \phi)$. The convergence constants $\sigma_{MC+MD}(E_0)$ and $\sigma_{MC}(E_0)$ are compared in Fig. 9: for a given number of points of the constant energy surface, the ensemble average $\langle A \rangle$ is computed with a greater accuracy if the Monte Carlo points are used to generate MD trajectories and time averages.

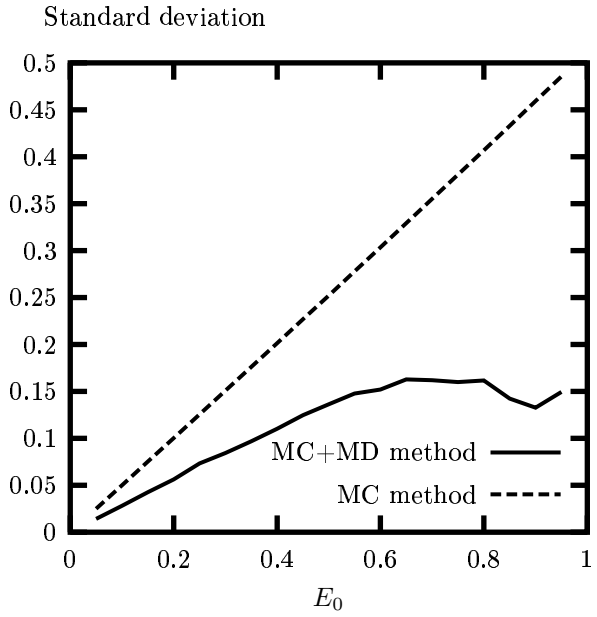

Figure 9: Computation of the ensemble average $\langle q_y^2 \rangle(E_0)$ on the double well potential: standard deviation $\sigma_{MC+MD}(E_0)$ when averaging MD time averages over initial conditions (solid line; we numerically checked that this quantity does not depend on the number of initial conditions), standard deviation $\sigma_{MC}(E_0)$ when using a standard MC method on Eq. (11) (dashed line).

Thus, to compute an ensemble average, we can suggest to choose several different initial conditions on the constant energy surface, and to use each of them to compute a trajectory and a time average. A good approximation of the ensemble average is the mean value of these time averages.

4 Conclusions

In order to compute ensemble averages of observables, one can use Molecular Dynamics and compute a time average of these observables over one (or several) trajectory of a dynamical

system. Under the ergodic assumption, this time average converges, as the trajectory length goes to infinity, towards the ensemble average. We have presented a numerical method to speed-up this convergence so that a trajectory of reduced length (and thus needing less computational effort) is needed to achieve a given precision. Along with this numerical method, we have presented sharp error bounds that may be used to optimize the simulation parameters.

A typical problem in Molecular Dynamics is the exploration of the whole phase space when the potential energy of the system has several basins corresponding to distinct metastable states. The method we have presented in this paper does not solve this issue (it does not improve the phase space exploration, but the computation of the time average), and other methods have to be used, for instance those suggested in Refs. [8, 10, 11].

ACKNOWLEDGMENTS

We would like to thank Gilles Zérah for his very helpful suggestions, Christian Lubich for the critical reading of the manuscript and Christophe Chipot for stimulating discussions. We also greatfully acknowledge the financial support of INRIA through the contract grant "Action de Recherche Concertée PRESTISSIMO".

5 Appendix

In this part, we detail the hypotheses needed to prove the estimates (4) and (7). To prove the estimate (9), an additionnal assumption is needed: the numerical scheme used to integrate the equations of motion should be symplectic. As we have mentioned it in Sec. 2, we consider a system of M particles in 3D, described by the Hamiltonian $H(\mathbf{q},\mathbf{p})$. Let $(\mathbf{q}_0,\mathbf{p}_0)$ be the initial condition.

5.1 Assumptions on the Hamiltonian function H:

We suppose that the Hamiltonian H is an analytical function, and that it is *completely integrable*, that is to say that (see Ref. [6], pp 214 and 272):

- there exists $3M$ invariant functions $I_j(\mathbf{q},\mathbf{p})$, $j=1,\ldots,3M$, of the dynamical system (1) (the definition of an invariant function is recalled in Sec. 2); we denote by $S(\mathbf{q},\mathbf{p})$ the level set of the invariant functions $\{I_j\}_{1\leq j\leq 3M}$ (cf. (2));
- these invariants I_j are in involution, i.e. satisfy the condition

  $$\forall (\mathbf{q},\mathbf{p})\in\mathbf{R}^{3M}\times\mathbf{R}^{3M},\ \ \forall j_1,j_2,\ \ \nabla_{\mathbf{q}} I_{j_1}(\mathbf{q},\mathbf{p})\cdot\nabla_{\mathbf{p}} I_{j_2}(\mathbf{q},\mathbf{p}) = \nabla_{\mathbf{p}} I_{j_1}(\mathbf{q},\mathbf{p})\cdot\nabla_{\mathbf{q}} I_{j_2}(\mathbf{q},\mathbf{p}),$$

  where $\nabla_{\mathbf{q}}$ is the gradient with respect to the position variables and $\nabla_{\mathbf{p}}$ is the gradient with respect to the momentum variables;

- there exists a neighbourhood of the initial condition $(\mathbf{q}_0, \mathbf{p}_0)$ such that, for all $(\mathbf{q}, \mathbf{p})$ in this neighbourhood, (i) the level set $S(\mathbf{q}, \mathbf{p})$ is compact and connected [12], and (ii) the gradients ∇I_j of the invariant functions are linearly independent;

5.2 Diophantine assumption:

Let $\mathbf{T}$ be the torus $\mathbf{R}/2\pi\mathbf{Z}$. Under the above hypotheses, it is possible [6, 7] to find a bounded open set $B \subset \mathbf{R}^{3M}$ containing $(I_1(\mathbf{q}_0, \mathbf{p}_0), \ldots, I_{3M}(\mathbf{q}_0, \mathbf{p}_0))$, a bounded open set $B' \subset \mathbf{R}^{3M} \times \mathbf{R}^{3M}$ containing the initial condition $(\mathbf{q}_0, \mathbf{p}_0)$ and a local change of variables ψ,

$$\psi : (\mathbf{a}, \theta) \in B \times \mathbf{T}^{3M} \quad \mapsto \quad (\mathbf{q}, \mathbf{p}) \in B' \subset \mathbf{R}^{3M} \times \mathbf{R}^{3M},$$

such that the new variables $(\mathbf{a}(t), \theta(t)) = \psi^{-1}(\mathbf{q}(t), \mathbf{p}(t))$ obey the simple dynamics

$$\frac{d\mathbf{a}}{dt} = 0, \qquad \frac{d\theta}{dt} = \omega(\mathbf{a}(t)). \tag{12}$$

supplied with initial condition $(\mathbf{a}(0), \theta(0)) = (\mathbf{a}_0, \theta_0) = \psi^{-1}(\mathbf{q}_0, \mathbf{p}_0)$. In (12), ω is a function from $\mathbf{R}^{3M}$ to $\mathbf{R}^{3M}$. The trajectory in the new variables is $\mathbf{a}(t) = \mathbf{a}_0$, $\theta(t) = \theta_0 + \omega(\mathbf{a}_0)t$.

We then make the following additional assumption (the so-called diophantine assumption) on the vector $\omega(\mathbf{a}_0)$: there exists a constant $C_0 > 0$ and an exponent $\gamma_0 \geq 0$ such that

$$\forall \alpha \in \mathbf{Z}^{3M} \setminus \{0\}, \quad \mid \alpha \cdot \omega(\mathbf{a}_0) \mid \geq \frac{C_0}{\mid \alpha \mid^{\gamma_0}},$$

where we note $\mid \alpha \mid = \alpha_1 + \ldots + \alpha_{3M}$ and $\alpha \cdot \omega(\mathbf{a}_0) = \alpha_1\omega_1(\mathbf{a}_0) + \ldots + \alpha_{3M}\omega_{3M}(\mathbf{a}_0)$. This hypothesis means that the quantity $\alpha \cdot \omega(\mathbf{a}_0)$ can get close to zero only if $\mid \alpha \mid$ goes to infinity.

Under the assumption that the Hamiltonian function H is analytic and integrable, and the diophantine assumption, one can prove the so-called ergodic theorem (see Sec. 2 and Refs. [5, 6]).

References

[1] M.P. Allen and D.J. Tildesley, *Computer simulation of liquids* (Oxford Science Publications, 1987).

[2] D. Frenkel and B. Smit, *Understanding molecular simulation, from algorithms to applications*, 2nd ed. (Academic Press, 2002).

[3] M.E. Tuckerman and G.J. Martyna, *Understanding modern molecular dynamics: techniques and applications*, J. Phys. Chem. B **104**, 159, 2000.

[4] S.D. Bond, B.J. Leimkuhler and B.B. Laird, *The Nosé-Poincaré method for constant temperature molecular dynamics*, J. Comput. Phys. **151**, 114, 1999.

[5] E. Cancès, F. Castella, Ph. Chartier, E. Faou, C. Le Bris, F. Legoll and G. Turinici, in preparation.

[6] V.I. Arnold, *Mathematical methods of classical mechanics* (Springer-Verlag, 1989).

[7] E. Hairer, Ch. Lubich, G. Wanner, *Geometric numerical integration, Structure-preserving algorithms for Ordinary Differential Equations* (Springer-Verlag, 2002).

[8] Ch. Schütte and W. Huisinga, in: *Handbook of numerical analysis, Vol. X, Special volume: Computational chemistry* (North-Holland, 2003).

[9] R.Y. Rubinstein, *Simulation and the Monte Carlo method* (Wiley, 1981).

[10] A. Laio, M. Parrinello, *Escaping free-energy minima*, Proc. Natl. Acad. Sci. USA **99**, 12562, 2002.

[11] M.R. Sorensen, A.F. Voter, *Temperature-accelerated dynamics for simulation of infrequent events*, J. Chem. Phys. **112**, 9599, 2000.

[12] G.F. Simmons, *Topology and modern analysis* (McGraw-Hill, 1963).

Contents

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399