Markov Chain Monte Carlo and Numerical Differential Equations

J.M. Sanz-Serna

1 Introduction

This contribution presents a —hopefully readable— introduction to Markov Chain Monte Carlo methods with particular emphasis on their combination with ideas from deterministic or stochastic numerical differential equations. Markov Chain Monte Carlo algorithms are widely used in many sciences, including physics, chemistry and statistics; their importance is comparable to those of the Gaussian Elimination or the Fast Fourier Transform. We have tried to keep the presentation as self-contained as it has been feasible. A basic knowledge of applied mathematics and probability[1] is assumed, but there are tutorial sections devoted to the necessary prerequisites in stochastic processes (Section 2), Markov chains (Section 3), stochastic differential equations (Section 6) and Hamiltonian dynamics/statistical physics (Section 8). The basic Random Walk Metropolis algorithm for discrete or continuous distributions is presented in Sections 4 and 5. Sections 7 and 9 are respectively devoted to MALA, an algorithm based on stochastic differential equations proposals, and to the Hybrid Monte Carlo method, founded on ideas from Hamiltonian mechanics.

We have avoided throughout mathematical technicalities (that in the study of continuous-time stochastic processes may be overwhelming). We have rather followed the style of presentation taken by D. Higham in his tutorial paper on stochastic differential equations [18] and aimed at an exposition based on computer experiments; we believe that this approach may provide much insight and be a very useful entry point to the study of the issues considered here.

J.M. Sanz-Serna
Depto. de Matemática Aplicada, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain. e-mail: sanzsern@mac.uva.es

[1] We assume notions such as discrete and continuous random variables, expectation, variance, conditional probability and independence.

2 Stochastic Processes

We begin with a few introductory definitions and some useful examples.

2.1 Preliminaries

The definition of stochastic process ([15], Chapter 8) is simple:

Definition 1. Let T be a set of indices. *A stochastic process* is a family $\{X_t\}_{t\in T}$ of random variables defined on a common probability space $(\Omega,\mathscr{A},\mathbb{P})$.[2]

In the applications, the variable t often corresponds to time. If $T=\mathbb{R}$, $T=[0,\infty)$ or $T=[a,b]$ the process is said to occur in *continuous time.* If $T=\{0,1,\dots\}$ the process takes place in *discrete time* and we write $\{X_n\}_{n\geq 0}$.

The variables X_t may take values in a *continuous state space* like $\mathbb{R}^d$ or in a finite or infinite *discrete state space* E. In the latter case and without loss of generality, one may assume that E has been identified with a subset of $\mathbb{Z}$.

By definition, X_t may be seen as a function of two arguments: t and ω.[3] For a given value of t, X_t is a function of ω (the chance), so that the value of X_t will be different in different instances of the random experiment. For a given draw of the chance ω, the value of X_t changes with time. In this way there are two different, complementary ways to study a given process $\{X_t\}_{t\in T}$:

- By studying the distribution of the variable X_t for each $t\in T$ and (since in all interesting cases the X_t's are not mutually independent) the distribution of the pair (X_{t_1},X_{t_2}) for each $t_1,t_2\in T$, ..., the distribution of $(X_{t_1},\dots,X_{t_n})$ for each $t_1,\dots,t_n\in T,\dots$
- By drawing ω from Ω and studying the map $t\mapsto X_t(\omega)$: *a trajectory or path or realization* of the process.

These considerations will hopefully become clearer with the examples that follow.

2.2 Some Simple Stochastic Processes

Let us examine three well-known, useful examples of time-discrete, discrete state space processes.

[2] Recall that (1) Ω is a set and each point $\omega\in\Omega$ corresponds to a possible outcome of a random experiment, (2) $\mathscr{A}$ (a σ-algebra) is the family of those subsets $A\subseteq\Omega$ called *events* to which a probability $\mathbb{P}(A)$ is assigned, (3) $\mathbb{P}$ is a probability measure, $\mathbb{P}:\mathscr{A}\to[0,1]$. The probability space plays very little explicit role in the study of the process; this is carried out in terms of the distributions of the X_t (see the examples in this section).

[3] Often the dependence of X_t on ω is not incorporated explicitly to the notation.

2.2.1 The Symmetric Random Walk

At each time $n = 1, 2, \ldots$, Mary and Paul toss a fair coin and bet one euro. Let X_n be Mary's accumulated gain before the $(n+1)$-st toss ($X_0 = 0$). Here the state space is $E = \mathbb{Z}$.

The distributions of the first few X_n are easily found:

$$\begin{aligned}
&\mathbb{P}(X_0 = 0) = 1;\\
&\mathbb{P}(X_1 = -1) = 1/2, \quad \mathbb{P}(X_1 = 1) = 1/2;\\
&\mathbb{P}(X_2 = -2) = 1/4, \quad \mathbb{P}(X_2 = 0) = 1/2, \quad \mathbb{P}(X_2 = 2) = 1/4;\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

The joint distribution of any pair (X_m, X_n) may also be determined readily. For instance for (X_3, X_4) we compute:

$$\begin{aligned}
&\mathbb{P}(X_3 = 1, X_4 = 0) = 3/16,\\
&\mathbb{P}(X_3 = 1, X_4 = 1) = 0,\\
&\mathbb{P}(X_3 = 0, X_4 = 1) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

For (X_1, X_2, X_3):

$$\begin{aligned}
&\mathbb{P}(X_1 = 1, X_2 = 2, X_3 = 1) = 1/8,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

Sets of four, five, ... X_n's are dealt with in the same way.

Note that $X_{n+1} = X_n + Z_n$,[4] where the variables Z_n (gain in toss $n+1$) are mutually independent and take values ± 1 with probability $1/2$ each. This leads to the formulae $\mathbb{E}(X_n) = 0$, $\mathrm{Var}(X_n) = n$ for the expectation and variance.

Fig. 1 shows, for $0 \le n \le 10$, two possible trajectories of the process. A computer-generated, longer trajectory may be seen in Fig. 2, where we note a few remarkable facts. (A complete study of the symmetric random walk using elementary means may be found in [9], Chapter 3.)

- The vertical axis covers only a small range slightly larger than $[-100, 100]$, in spite of the fact that Mary's gains might in principle have been in the range $-10,000 \le X_n \le 10,000$. This happens because the standard deviation $\sigma(X_n)$ equals $\sqrt{n}$.
- While the game is 'fair' i.e. $\mathbb{E}(X_n) = 0$, Mary has been winning most of the time. This is not a peculiarity of the particular trajectory shown: typically either Mary is ahead most of the time or Paul is ahead most of the time.

[4] In general, a process $\{X_n\}_{n\ge 0}$ is a *random walk* if $X_{n+1} = X_n + Z_n$, where Z_n is independent of $X_n, \ldots, X_0$.

- In the first two or three thousand tosses Mary and Paul tied ($X_n = 0$) a few times. Since after a tie the game restarts afresh —the coin has no memory— one would have expected that similar ties would keep happening after, say, $n = 3,000$. Clearly this has not been the case. Our intuition suggests that, if T_i is the number of tosses between consecutive ties, then the average $A_n = (T_1 + \cdots + T_n)/n$ should converge to a limit as $n \to \infty$. However it may be proved that the size of A_n grows proportionally to n, so that $T_1 + \cdots + T_n$ grows like n^2. In fact it is likely that one among $T_1, \ldots, T_n$ be of size proportional to n^2.

2.2.2 The Non-symmetric Random Walk

Everything is as before but Mary's chance of winning an individual bet is now $p \neq 1/2$ so that Paul's is $q = 1 - p$.

For the distributions of the X_n we find now:

$$\begin{aligned}
&\mathbb{P}(X_0 = 0) = 1,\\
&\mathbb{P}(X_1 = -1) = q, \quad \mathbb{P}(X_1 = 1) = p,\\
&\mathbb{P}(X_2 = -2) = q^2, \quad \mathbb{P}(X_2 = 0) = 2pq, \quad \mathbb{P}(X_2 = 2) = p^2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

From $X_{n+1} = X_n + Z_n$ we compute $\mathbb{E}(X_n) = n(p-q)$ and $\mathrm{Var}(X_n) = 4npq$. Since the expectation grows like n and the standard deviation only like $\sqrt{n}$, the *drift* arising from the lack of fairness of the coin, $p \neq q$, will in the long run dominate the *fluctuations* due to chance, even if $|p-q|$ is very small. This is borne out in Fig. 3, where $p = 0.55$.

2.2.3 The Ehrenfest Diffusion Model

This was proposed in 1907 by P. Ehrenfest (1880–1933) to illustrate the second law of thermodynamics. Two containers, left and right, are adjacent to each other and contain gas that may move between them through a small aperture. There are in total M molecules. At each time n, a molecule, randomly chosen among the M, moves to the other container. Let X_n be the number of molecules in the left box before the $(n+1)$-st move. We assume that initially all molecules are in the left container, $X_0 = M$. The state space is $\{0, 1, \ldots, M\}$ and the distributions are:

$$\begin{aligned}
&\mathbb{P}(X_1 = M-1) = 1,\\
&\mathbb{P}(X_2 = M-2) = (M-1)/M, \quad \mathbb{P}(X_2 = M) = 1/M,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

A typical trajectory may be seen in Fig. 4, where we observe that, after an initial transient, X_n has not left the interval $[25, 75]$, i.e. the molecules distribute themselves

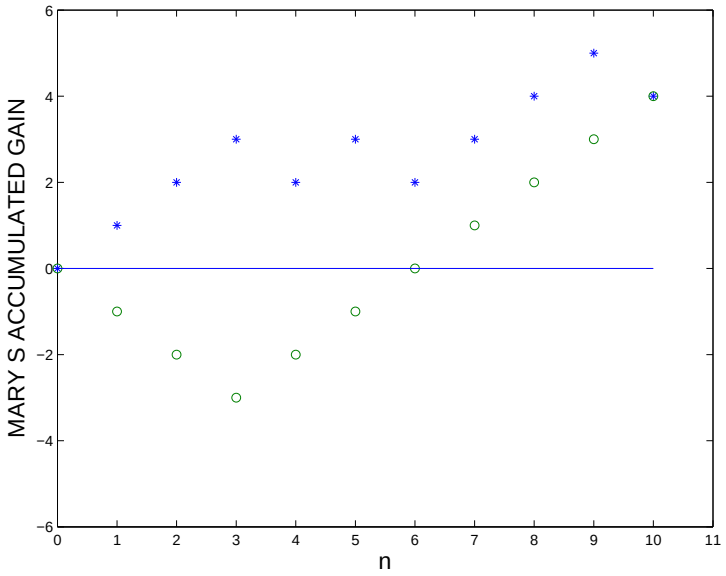

Fig. 1 Two possible trajectories of the symmetric random walk, $0 \leq n \leq 10$

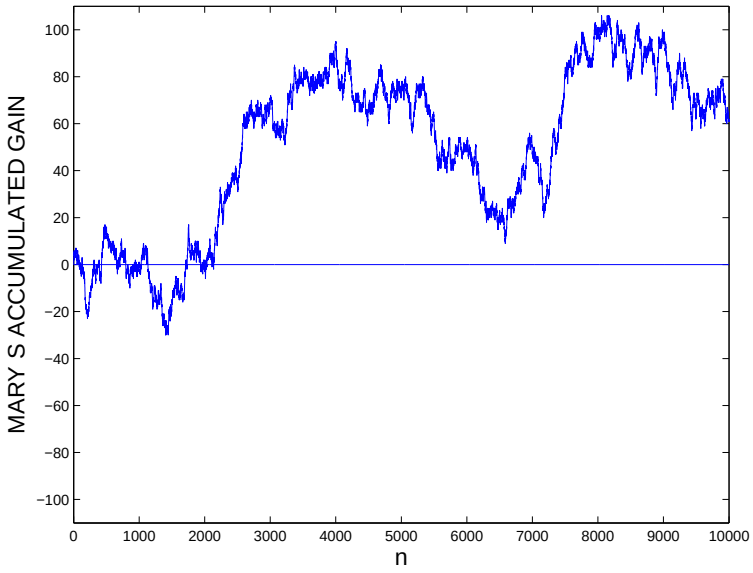

Fig. 2 A typical trajectory $0 \leq n \leq 10,000$ of the symmetric random walk

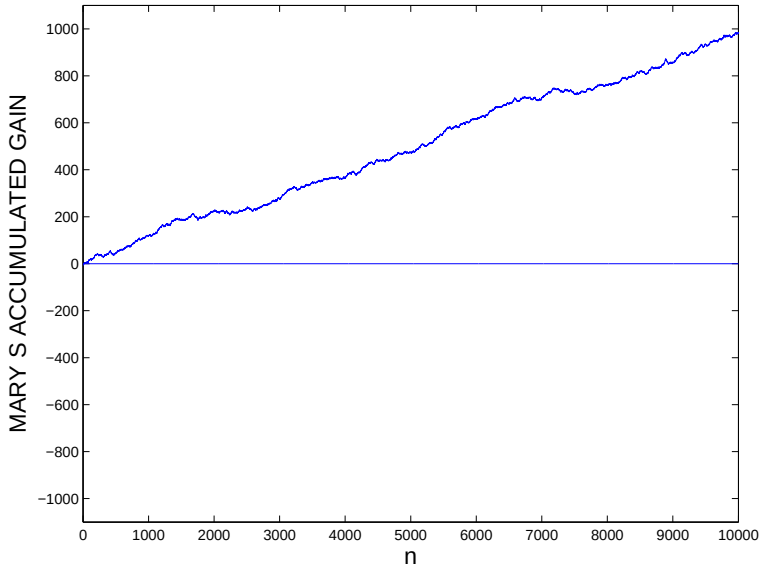

Fig. 3 Non-symmetric random walk, $p = 0.55$, $0 \leq n \leq 10,000$. The expectation of $X_{10,000}$ is 1,000 and its typical deviation ≈ 100. Drift offsets fluctuations due to chance

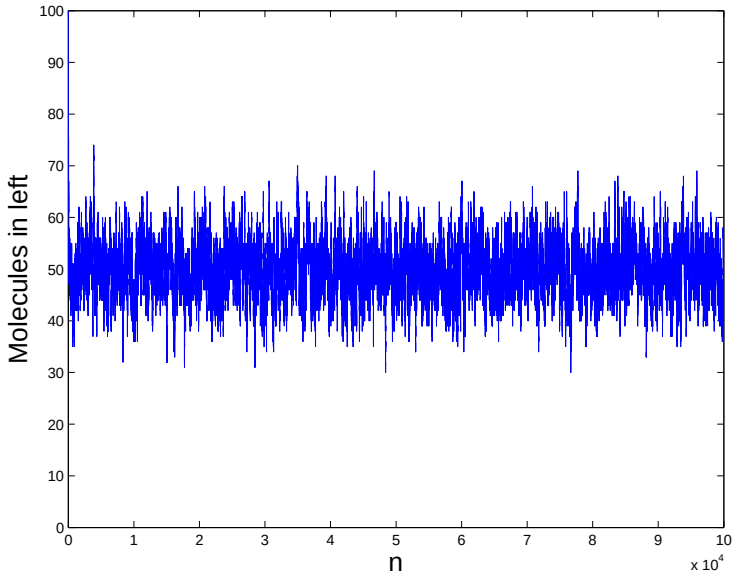

Fig. 4 A trajectory of the Ehrenfest model, 100,000 steps for $M = 100$ molecules. Outside the initial transient X_n has not left the interval $[25, 75]$

more or less evenly between both containers. It may be shown ([9], Chapter XVII, Example 7.c) that, with $M = 10^6$ the probability of finding, for n outside the initial transient, more than $505,000$ molecules in one container (i.e. of finding fluctuations larger than one percent around the break-even situation $X_n = M/2$) is of the order of 10^{-23}. In statistical physics M is of course much, much larger and the size of the fluctuations around $M/2$ correspondingly smaller: for all practical purposes the gas, driven by sheer chance, will remain in the maximum entropy state $X_n = M/2$.

3 Discrete State Space Markov Chains

A Markov chain (MC) is a process $\{X_n\}_{n\geq 0}$ where the distribution of X_{n+1} conditional on $X_0,\ldots,X_n$ coincides with the distribution of X_{n+1} conditional on X_n. This is sometimes expressed by saying 'in order to know the future, the knowledge of the past does not add anything to the knowledge of the present'. The precise definition is:

Definition 2. A discrete-time process $\{X_n\}_{n\geq 0}$ with values in a countable space E is a *Markov chain* if for all $n \geq 0$ and all states $i_0, i_1, \ldots, i_{n-1}, i, j \in E$

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

(whenever both sides are well defined[5]).

Markov chains are named after A. Markov (1856–1922). Cases with infinite, countable state spaces where first considered by A. Kolmogorov in 1936.

If, for each pair of states $i, j \in E, \mathbb{P}(X_{n+1} = j \mid X_n = i)$ is independent of n, then the MC is called *homogeneous*; only homogeneous MC are considered in this paper.

The symmetric and non-symmetric random walks in Section 2.2.1, 2.2.2 are (homogeneous) MC: the structure $X_{n+1} = X_n + Z_n$ noted before makes it clear that the knowledge of the values of $X_0,\ldots,X_{n-1}$ adds nothing to the knowledge of the value of X_n (in both trajectories in Fig. 1 $X_{10} = 4$ and the distribution of X_{11} conditional on the past is the same: $X_{11} = 5$ or $X_{11} = 3$ with probability 1/2 each). In the symmetric random walk, the so-called transition probabilities are

$$\begin{aligned}
&\mathbb{P}(X_{n+1} = i+1 \mid X_n = i) = \mathbb{P}(Z_n = 1) = 1/2,\\
&\mathbb{P}(X_{n+1} = i-1 \mid X_n = i) = \mathbb{P}(Z_n = -1) = 1/2,\\
&\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(Z_n \neq \pm 1) = 0, \quad j \neq i \pm 1.
\end{aligned}$$

The Ehrenfest model (Section 2.2.3) is another example of MC. The transition probabilities are:

[5] Note that e.g. $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ does not make sense if $\mathbb{P}(X_n = i) = 0$. Here we shall not pay attention to the difficulties created by probabilities conditioned to events $\{X_n = i\}$ of 0 probability. These difficulties are easily avoided if, as in [9], the Markov chain is defined in the first place by means of the transition probabilites rather than in terms of the variables X_n.

$$\mathbb{P}(X_{n+1} = i+1 \mid X_n = i) = (M-i)/M, \quad i < M,$$
$$\mathbb{P}(X_{n+1} = i-1 \mid X_n = i) = i/M, \quad i > 0$$

(other transitions are impossible).

3.1 The Transition Matrix

The *transition probabilities* of a MC are defined by

$$p_{ij} := \mathbb{P}(X_{n+1} = j \mid X_n = i).$$

Since, as j ranges in E with i fixed, the p_{ij} are a probability distribution, we may write

$$p_{ij} \geq 0, \qquad \sum_{j \in E} p_{ij} = 1,$$

so that the *transition matrix* $P = (p_{ij})_{(i,j)\in E\times E}$ is a *stochastic matrix*.[6] The i-th row of P provides the distribution of X_{n+1} conditional on $X_n = i$.

A transition over two steps from i to k must be accomplished through an intermediate visit to some state j and therefore we may write the *Chapman-Kolmogorov equation*

$$\mathbb{P}(X_{n+2} = k \mid X_n = i) = \sum_j \mathbb{P}(X_{n+1} = j \mid X_n = i)\, \mathbb{P}(X_{n+2} = k \mid X_{n+1} = j) = \sum_j p_{ij} p_{jk}.$$

In this way $\mathbb{P}(X_{n+2} = k \mid X_n = i)$ is given by the (i,k) entry of P^2. The entries of higher powers P^3, P^4, ..., give similarly the probabilities of transitions in $3, 4, \ldots$ steps.

The (unconditional) distribution of each X_n is determined by the distribution of X_0 together with the transition matrix P:

$$\mathbb{P}(X_n = \ell) = \sum_i \mathbb{P}(X_0 = i)\, \mathbb{P}(X_n = \ell \mid X_0 = i) = \sum_i \mathbb{P}(X_0 = i)\, (P^n)_{i\ell}.$$

It is customary to collect in a column vector $\mu^{(n)}$ the probabilities $\mathbb{P}(X_n = \ell)$, $\ell \in E$, and then the preceding formula may be rewritten as

$$\mu^{(n)T} = \mu^{(0)T} P^n.$$

Equivalently one has the following expression for the evolution of the distributions $\mu^{(n)}$:

$$\mu^{(n+1)T} = \mu^{(n)T} P, \qquad n = 0, 1, \ldots \tag{1}$$

[6] If E comprises an infinite number of states, this 'matrix' will of course have infinitely many rows/columns. Sums like $\sum_j p_{ij} p_{jk}$ that we shall find below have a finite value if P is stochastic.

The joint distributions of (X_m, X_n), (X_ℓ, X_m, X_n), ... are also easily determined once $\mu^{(0)}$ and P are known. In practice it is customary to describe a MC by specifying the transition matrix together with the initial distribution.[7] Although strictly speaking the MC is the sequence $\{X_n\}$, in practice we often speak as though the chain were the matrix P together with the initial distribution $\mu^{(0)}$ or even the matrix P with an undetermined $\mu^{(0)}$.

3.2 Classifying Markov Chains

The following concept is needed in order to define persistent states:

Definition 3. For each state $i \in E$ define the *return time* T_i (a random variable) by:

$$T_i := \inf\{n \geq 1 : X_n = i\} \in [0, \infty].$$

Here it is understood that the inf of the empty set is ∞. Note that $n \geq 1$ so that, in particular $X_0 = i$ does not imply $T_i = 0$.

Definition 4. A state $i \in E$ is *persistent or recurrent* if $\mathbb{P}(T_i < \infty \mid X_0 = i) = 1$. Otherwise it is called transient.

A recurrent state $i \in E$ is *positive* if $\mathbb{E}(T_i \mid X_0 = i) < \infty$. Otherwise it is called *null*.

If i is persistent and the chain is started at i, then the number of visits to i (i.e. the number of values of n with $X_n = i$) is infinite with probability 1 (see [5] Theorem 8.2).

For the chain

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix} \tag{2}$$

the second state is persistent. In fact, if started at 2, the chain returns to 2 in one move with probability 1/2, in two moves with probability 1/4, etc. The third state is persistent for the same reason. The first state is transient: conditional to $X_0 = 1$, T_1 only takes the values 1 (with probability $1/2$) and ∞.

The matrix (2) is certainly special: states 2 and 3, on their own, would make up a MC. This matrix provides an example of reducibility in the sense of the next definition because moves from 2 to 1 or from 3 to 1 in n steps are impossible for each $n \geq 1$.

Definition 5. A MC is *irreducible* if for each ordered pair of states $i, j \in E$ there exists a number $n = 1, 2, \ldots$ such that

$$\mathbb{P}(X_n = j \mid X_0 = i) > 0.$$

[7] See [5], Theorem 8.1 for the construction of the X_n and the underlying probability space.

The following important result holds ([6] Chapter 3, Section 1.3):

Theorem 1. *For an irreducible chain, one of the following three possibilities holds true:*

- *All states are positive recurrent.*
- *All states are null recurrent.*
- *All states are transient.*

If, in addition, E is finite, then the chain is necessarily positive recurrent.

The Ehrenfest model, the symmetric random walk, and the non-symmetric random walk are irreducible and provide examples of the three possibilities in the theorem ([6] Chapter 3, Example 1.2). For the symmetric random walk the fact that the expected waiting time for the first tie $X_n = 0$ is infinite is related to some of the counterintuitive features we saw in Fig. 2; for instance to the fact that the average A_n of the first n times between successive returns to equilibrium does not approach a finite limit. For the Ehrenfest model positive recurrence implies that, except for a set of trajectories with probability 0, in each trajectory there are infinitely many n_r such that $X_{n_r} = M$: all the molecules will be back in the left container infinitely many times! There is no contradiction with Fig. 4: the expectation $\mathbb{E}(T_M \mid X_0 = M)$ is positive but exponentially small as we shall see.

3.3 Stationary Distributions

The following concept is extremely important.

Definition 6. A probability distribution μ on E ($\mu_i \geq 0$, $i \in E$, $\sum_i \mu_i = 1$) is called a *stationary or invariant or equilibrium* distribution of the MC with transition matrix P if

$$\sum_i \mu_i p_{ij} = \mu_j, \qquad j \in E,$$

or, in matrix notation,

$$\mu^T P = \mu^T. \tag{3}$$

From (1) it follows that if X_0 possesses the distribution μ and μ is invariant, then all the X_n, $n = 0, 1, 2, \ldots$ share the same distribution; we then say that the chain is at *stationarity*. Note that at stationarity the X_n are identically distributed but, except for trivial cases,[8] *not independent*. It is easy to see that the symmetric and non-symmetric random walk do not possess invariant probability distributions. For the Ehrenfest model the distribution

$$\mu_i = \binom{M}{i} \frac{1}{2^M} \tag{4}$$

[8] If all the rows of P are equal, X_{n+1} is independent of X_n.

is readily seen to be invariant. Note that μ_i coincides with the probability that, when each of the M molecules is randomly assigned to the left or right container (with probability $1/2$ each), then the left container receives i molecules.

As a further example consider the *doubly stochastic matrix* ($\sum_j p_{ij} = \sum_i p_{ij} = 1$, $p_{ij} \geq 0$):

$$P = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \\ 1/2 & 1/4 & 1/4 \end{bmatrix}. \tag{5}$$

The invariant distribution is $[1/3, 1/3, 1/3]$: at stationarity all states have the same probability, something that happens with all doubly stochastic transition matrices.

The following general result holds ([6] Chapter 3, Theorems 3.1, 3.2):

Theorem 2. *Assume the chain to be irreducible. Then it is positive recurrent if and only if there is a stationary distribution. The stationary distribution μ, if it exists is unique and $\mu_i = 1/\mathbb{E}(T_i \mid X_0 = i) > 0$ for each $i \in E$.*

Since from (4) $\mu_M = 2^{-M}$, the theorem shows that in Fig. 4 the expected number of interchanges for all 100 molecules to return to the left container is $2^{100} \approx 1.27 \times 10^{30}$.

3.4 Reversibility

Consider a MC with a stationary distribution such that $\mu_i > 0$ for each $i \in E$ (a particular case is given by an irreducible, positive recurrent chain, see Theorem 2). The matrix Q with entries $q_{ij} = \mu_j p_{ji}/\mu_i$ is stochastic,

$$\sum_j q_{ij} = \sum_j \frac{\mu_j p_{ji}}{\mu_i} = \frac{\mu_i}{\mu_i} = 1,$$

and also has μ_i as an invariant distribution

$$\sum_i \mu_i q_{ij} = \sum_i \mu_j p_{ji} = \mu_j \sum_i p_{ji} = \mu_j.$$

What is the meaning of Q? Assume that the initial distribution $\mu^{(0)}$ coincides with μ (i.e. the chain is at stationarity) then:

$$\mathbb{P}(X_n = j \mid X_{n+1} = i) = \frac{\mathbb{P}(X_{n+1} = i \mid X_n = j)\,\mathbb{P}(X_n = j)}{\mathbb{P}(X_{n+1} = i)} = \frac{p_{ji}\,\mu_j}{\mu_i} = q_{ij}.$$

Thus, Q is the transition matrix of a chain where the 'arrow of time' has been reversed, because n and $n+1$ have interchanged their roles. As a simple example consider the chain (5) for which $Q = P^T$, since the stationary distribution is $\mu_1 = \mu_2 = \mu_3 = 1/3$. If P is at stationarity the three events $\{X_n = 1, X_{n+1} = 2\}$, $\{X_n = 2, X_{n+1} = 3\}$, $\{X_n = 3, X_{n+1} = 1\}$ have probability $1/6$ each, while the six

events $\{X_n = 1, X_{n+1} = 1\}$, $\{X_n = 1, X_{n+1} = 3\}$, etc. have probability $1/12$ each. For Q it is the events $\{X_n = 2, X_{n+1} = 1\}$, $\{X_n = 3, X_{n+1} = 2\}$, $\{X_n = 1, X_{n+1} = 1/3\}$ that have probability $1/6$.

Definition 7. A probability distribution $\mu > 0$ and a transition matrix P satisfy the *detailed balance* condition if:

$$\forall i, j \in E, \qquad \mu_i p_{ij} = \mu_j p_{ji}. \tag{6}$$

Of course $\mu_i p_{ij}$ is the probability of the event $(X_n = i, X_{n+1} = j)$. For the example (5) the detailed balance condition does not hold: the event $(X_n = 1, X_{n+1} = 2)$ is more likely than the event $(X_n = 2, X_{n+1} = 1)$.

Since (6) implies

$$\sum_i \mu_i p_{ij} = \sum_i \mu_j p_{ji} = \mu_j,$$

we may conclude:

Theorem 3. *Under the assumption of detailed balance (6):*

- *μ is an invariant distribution with respect to P.*
- *At stationarity, the reversed matrix Q coincides with P and therefore the chain and its time-reversal are statistically the same.*

The chain is then called reversible *with respect to μ.*

The Ehrenfest chain is in detailed balance with the distribution (4) and hence reversible with respect to it.

3.5 Ergodicity

The ergodic theorem ([6] Chapter 3, Theorem 4.1) is the foundation of all our later work:

Theorem 4. *Let $\{X_n\}_{n\geq 0}$ be an irreducible, positive recurrent MC and denote by μ its stationary distribution as in Theorem 2. For any function $f : E \to \mathbb{R}$ such that*

$$\sum_{i\in E} |f(i)|\mu_i < \infty$$

and any initial distribution $\pi^{(0)}$, $\mathbb{P}_{\mu^{(0)}}$ almost sure:

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} f(X_k) = \sum_{i\in E} f(i)\mu_i. \tag{7}$$

The last sum is of course the expectation of $f(X)$ when X has the distribution μ. Fig. 5 shows the ergodic theorem at work in Ehrenfest's model. The $+$ signs

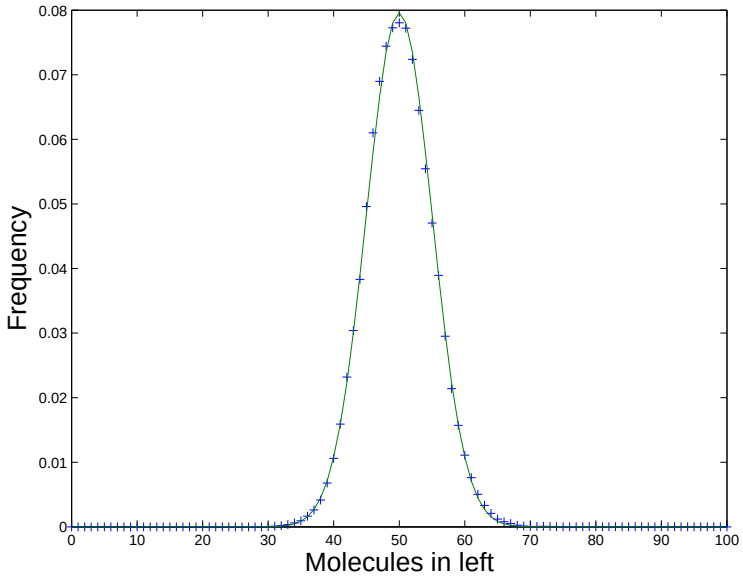

Fig. 5 Ehrenfest's model: an histogram of the trajectory in Fig. 4 (+ signs) and the invariant distribution (4) (solid line)

measure the frequency with which the $M+1$ states $0,\ldots,M$ have been occupied along the single trajectory of Fig. 4 and the solid line represents the equilibrium probabilities (4). For each state i, the probability in (4) is of course the expectation of the function f such that $f(i)=1$ and $f(j)=0$ for $j\neq i$ (the indicator of the set $\{i\}$) and the frequency of occupation is the average of f along the trajectory; we see in Fig. 5 that both very approximately coincide, in agreement with the ergodic theorem. In statistical physics one considers an *ensemble*, i.e. a very large number of Ehrenfest experiments running in parallel, independently of one another, so that, at any fixed n, the distribution of the number of molecules in the left containers across the experiments coincides with the distribution of the random variable X_n. The ergodic theorem implies that the behavior of a trajectory in a single experiment started e.g. from $X_0=M$ coincides with the behavior of the ensemble at any fixed time n *when the initial distribution across the ensemble is given by (4), i.e. if the chain is at stationarity*.

Ergodicity makes it possible to compute expectations with respect to the stationary distribution by computing averages along trajectories. This is the basis of Monte Carlo algorithms that we shall study in the next sections.

3.6 Convergence to Steady State

With $M = 2$ molecules, the Ehrenfest transition matrix is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix};$$

it is easy to check that returns to the initial state $X_n = X_0$ are only possible if n is even. This motivates the following definition.

Definition 8. The period $d(i)$ of a state i is the greatest common divisor of all the numbers n such that a return to i in n steps is possible. If $d(i) = 1$, the state i is said to be *aperiodic*. If $d(i) > 1$ the state i is said to be *periodic* with period $d(i)$.

It turns out that for an irreducible chain either all states are aperiodic or all states are periodic and share the same period ([15] Section 6.3). One then says that the chain is aperiodic or periodic respectively. The Ehrenfest chain, regardless of the number of molecules M, is periodic with period 2 and so are the symmetric and non-symmetric random walks of Sections 2.2.1 and 2.2.2.[9]

The result below ([6], Theorem 2.1) shows that, if we exclude periodic chains, the distribution $\mu^{(n)}$ of X_n in an irreducible, positive recurrent chain will approach as $n \uparrow \infty$ the stationary distribution, regardless of the choice of the distribution of X_0. This fact is sometimes expressed by saying that an irreducible, aperiodic and positive recurrent chain is *asymptotically stationary*; for n large the chain 'forgets its origin.'

Theorem 5. *Assume that a MC is irreducible, positive recurrent and aperiodic and let μ be the corresponding invariant distribution as in Theorem 2. For each choice of the distribution*[10] *$\mu^{(0)}$ of X_0*

$$\lim_{n\to\infty} \mid \mu^{(0)T} P^n - \mu^T \mid = \lim_{n\to\infty} \mid \mu^{(n)T} - \mu^T \mid = 0.$$

It is perhaps useful to stress that the ergodic Theorem 4 holds for both periodic and aperiodic (irreducible, positive recurrent) chains (after all Fig. 5 illustrates this theorem at work in the periodic Ehrenfest chain). However, if the chain is irreducible, positive recurrent and aperiodic, the combination of Theorems 4 and 5 implies that the average in the left-hand side of (7) approximately coincides with the expectation of f with respect to the distribution $\mu^{(n)}$ of X_n, $n \gg 1$, *regardless of the choice of the distribution* $\mu^{(0)}$.

---

[9] This should not lead to the conclusion that period 2 is the rule for MCs. The three examples in the last section are not typical in this respect and were chosen in view of the fact that are very easily described—in each of them transitions may only occur between state i and states $i \pm 1$—. Any chain where the diagonal elements of P are all $\neq 0$ only contains aperiodic states.

[10] If λ, ν are distributions the notation $\mid \lambda^T - \nu^T \mid$ means $\sum_i |\lambda_i - \nu_i|$.

In this connection consider an ensemble of many couples, Marys and Pauls, with each couple tossing a coin and betting repeatedly ($X_0 = 0$).[11] At any fixed n, the distribution of X_n is symmetric and there will be approximately as many Marys ahead as Pauls ahead —this is a property of the *ensemble*—. However, as illustrated in Fig. 2, in the typical trajectory of a single couple either Mary or Paul will be ahead most of the time. Here what is typical for a single trajectory is markedly different from what happens in the ensemble; this difference explains why some people find Fig. 2 disconcerting.

4 Sampling from a Target Distribution by Monte Carlo Algorithms: The Discrete Case

Markov Chain Monte Carlo (MCMC) algorithms [28] are aimed at computing expectations[12] with respect to a given *target distribution* μ on a state space E, that for the time being we assume discrete. These algorithms construct a MC $\{X_n\}_{n\geq 0}$ for which μ is an invariant distribution and, as pointed out before, invoke ergodicity to approximate expectations by empirical means along a trajectory as in (7).

MCMC is a very popular technique in computational chemistry and physics; examples will be considered in later. Bayesian statistics ([30] Section 1.3) is another important field of application. There a *prior* probability distribution μ_0 is postulated on the (discrete) set Θ of possible values of a parameter θ appearing in a probabilistic model (note that Θ is now the state space and that the different states are the different values of θ). Then data y are collected and 'incorporated' into the model via Bayes's theorem to define a *posterior* distribution for θ

$$\mu(\theta \mid y) = \frac{\phi(y \mid \theta)\mu_0(\theta)}{\sum_{\zeta\in\Theta} \phi(y \mid \zeta)\mu_0(\zeta)} \tag{8}$$

($\phi(y \mid \theta)$ is the probability of y when the parameter value is θ.) As a rule, the posterior is neither one of the familiar distributions nor tractable analytically and in order to compute expectations

$$\mathbb{E}(f(\theta) \mid y) = \sum_{\theta\in\Theta} f(\theta)\mu(\theta \mid y)$$

it is necessary to resort to Monte Carlo techniques.

The idea behind MCMC methods [26] was suggested by Metropolis and his coworkers in 1953. The problem to be addressed is essentially[13] how to construct a

[11] Note that there is no invariant probability distribution, since the chain is null recurrent.

[12] Of course computing the probability of an event A is equivalent to computing the expectation of its indicator, i.e. of the random variable that takes the value 1 if $\omega \in A$ and 0 if $\omega \notin A$.

[13] It is also necessary that the chain constructed be positive recurrent. Also not all positive recurrent chains having the target as equilibrium measure are equally efficient, as the velocity of the convergence to the limit in (7) is of course chain-dependent.

MC $\{X_n\}_{n\geq 0}$ in a given state space E having the target μ as an invariant distribution or, more precisely, how to compute trajectories of that chain in order to be able to use the empirical average in the left-hand side of (7) as an approximation to the expectation in the right-hand side. Metropolis algorithms use two ingredients to obtain realizations from $\{X_n\}_{n\geq 0}$:

1. Realizations u_0, u_1 ... of a sequence of mutually independent random variables $U_0, U_1, \ldots$ with uniform distribution in the unit interval. These realizations are of course readily available on any computing system.
2. Samples from the distribution of Y_{n+1} conditional on Y_n in an *auxiliary* MC $\{Y_n\}_{n\geq 0}$ in the same state space E (not in the MC $\{X_n\}_{n\geq 0}$ we wish to construct!). At this stage the only requirement we impose on $\{Y_n\}_{n\geq 0}$ is that the transition probabilities p^*_{ij} satisfy the symmetry requirement

$$\forall i, j, \qquad p^*_{ij} = p^*_{ji}. \tag{9}$$

For example, if $E = \mathbb{Z}$ then $\{Y_n\}_{n\geq 0}$ may be defined through $Y_{n+1} = Y_n + Z_n$, where the Z_n are mutually independent, integer-valued and with a symmetric distribution (i.e. $\mathbb{P}(Z_n = i) = \mathbb{P}(Z_n = -i)$ for each $i \in \mathbb{Z}$). To generate a sample of $Y_{n+1} \mid Y_n = y_n$ just set $y_{n+1} = y_n + z_n$ where z_n is a sample of Z_n. If $Z_n = \pm 1$ with probability $1/2$ each, then $\{Y_n\}_{n\geq 0}$ is of course the symmetric random walk in Section 2.2.1.

The algorithm is as follows:

- Choose a value i_0 for X_0 (randomly or e.g. $i_0 = 0$).
- Once values $i_0, \ldots, i_n$ of $X_0, \ldots, X_n$ have been found:
  - Generate a *proposed* value $i^*_{n+1} \in E$, from the auxiliary conditional distribution $Y_{n+1} \mid Y_n = i_n$.
  - If $\mu_{i^*_{n+1}}/\mu_{i_n} > u_n$ set $X_{n+1} = i^*_{n+1}$; in this case we say that *the proposal is accepted*. Else set $X_{n+1} = i_n$ and we say that *the proposal is rejected*.

The criterion used to accept or reject the proposal is called *the Metropolis accept/reject rule*. After noting that the acceptance probability is[14]

$$a = 1 \wedge \frac{\mu_{i^*_{n+1}}}{\mu_{i_n}} \tag{10}$$

(i.e. the proposal is accepted with probability a), it is not difficult to prove the following result:

Theorem 6. *The transitions $X_n \mapsto X_{n+1}$ in the procedure just described satisfy the detailed balance condition (6) with respect to the target distribution μ. Therefore (Theorem 3), the implied chain $\{X_n\}_{n\geq 0}$ is reversible with respect to μ.*

Proof. If $j \neq i$, to reach j at step $n+1$ we require (i) that j is proposed and (ii) that it is accepted. In this way:

[14] $\wedge$ means min.

$$\mu_i\, p_{ij} = \mu_i \left(p^*_{ij}\, (1 \wedge \frac{\mu_j}{\mu_i}) \right) = p^*_{ij}\, (\mu_i \wedge \mu_j),$$

and the last expression is symmetric in i, j in view of (9). □

A few remarks:

- The target distribution μ only enters the algorithm through the ratios in (10) and hence must be known only up to a multiplicative constant. This is an asset: in many applications the normalizing constant of the target distribution is extremely difficult to determine. As an example look at (8) where, for given values of the data y, the denominator is just a real number that normalizes the distribution (i.e. it ensures that, as θ ranges in Θ, the values of $\mu(\theta \mid y)$ add up to 1). In practice, the computation of that denominator may be impossible if the cardinality of Θ is large; when computing the acceptance ratio in the Metropolis algorithm one may substitute the un-normalized values $\phi(y \mid \theta)\mu_0(\theta)$ for the true probabilities $\mu(\theta \mid y)$.
- The rejected values of X_n are part of the chain and must be included to compute the average
$$\frac{1}{N+1} \sum_{k=0}^{N} f(X_k)$$
used to approximate the expectation of f.
- In practice if the starting location i_0 of the chain is far away from the states i for which the target has a significant size the convergence in (7) will be very slow. It is then advisable to run the chain for a *burn in* preliminary period until the states of higher probability are identified. The values of X_n corresponding to the burn in phase are not used to compute the averages.

As we shall see later, it is of interest to consider proposals that do not satisfy the symmetry condition in (9). (For instance we may wish to use proposals $Y_{n+1} = Y_n + Z_n$ where the increments Z_n do not have a symmetric distribution.) As first pointed out by Hastings in 1970 [17], to achieve detailed balance the formula (10) for acceptance probability has then to be changed into

$$a = 1 \wedge \frac{p^*_{i^*_{n+1} i_n}\, \mu_{i^*_{n+1}}}{p^*_{i_n i^*_{n+1}}\, \mu_{i_n}}. \tag{11}$$

The proof that this so-called Metropolis-Hastings rule works follows the lines of the proof of Theorem 6. Further possibilities for the acceptance probability recipe exist [17].

5 Metropolis Algorithms for the Continuous Case

For the sake of simplicity, our presentation of the Metropolis-Hastings algorithms has assumed that the target probability is defined on a discrete state space. However the algorithms are equally applicable to sampling from continuous distributions and in fact the next sections will only deal with the continuous case. We begin with a few words on MC with a continuous state space.

5.1 Continuous State Space Markov Chains

We now consider (time-discrete) stochastic processes $\{X_n\}_{n\geq 0}$ where each X_n takes values in $\mathbb{R}^d$. The definition of MC remains the same: $\{X_n\}_{n\geq 0}$ is a MC if the distribution of X_{n+1} conditional on X_n and the distribution of X_{n+1} conditional on $X_n,\ldots,$ X_0 coincide. The role played in the discrete state space case by the transition matrix P is now played by a *transition kernel* K (see e.g. [30] Definition 6.2, [10] Chapter VI, Section 11). This is a real-valued function $K(\cdot,\cdot)$ of two arguments. For each fixed $x\in\mathbb{R}^d$, $K(x,\cdot)$ is a Borel probability measure in $\mathbb{R}^d$. For each fixed Borel set $A\subseteq\mathbb{R}^d$, $K(\cdot,A)$ is a Borel measurable function. The value $K(x,A)$ represents the probability of jumping from the point $x\in\mathbb{R}^d$ to a set A in one step of the chain. Hence the formula (1) for the evolution of the distributions $\mu^{(n)}$ of the X_n now becomes

$$\mu^{(n+1)}(A)=\mathbb{P}(X_{n+1}\in A)=\int_{\mathbb{R}^d}\mu^{(n)}(dx)\,K(x,A),$$

or in shorthand

$$\mu^{(n+1)}(dy)=\int_{\mathbb{R}^d}\mu^{(n)}(dx)\,K(x,dy). \tag{12}$$

The condition (3) for a stationary or invariant probability distribution is correspondingly

$$\mu(A)=\int_{\mathbb{R}^d}\mu(dx)\,K(x,A)$$

(for each measurable A) or

$$\mu(dy)=\int_{\mathbb{R}^d}\mu(dx)\,K(x,dy), \tag{13}$$

and the detailed balance condition (6) reads

$$\int_A\mu(dx)\,K(x,B)=\int_B\mu(dy)\,K(y,A), \tag{14}$$

or

$$\mu(dx)K(x,dy)=\mu(dy)K(y,dx).$$

While conditions exist that guarantee the existence of a stationary probability distribution and the validity of a corresponding ergodic theorem (see [27], [30] Chapter

6) the technicalities are much more intricate than in the discrete state space case and will not be studied here.

In practice the kernel K often possesses a density $k(x,y)$ (with respect to the standard Lebesgue measure in $\mathbb{R}^d$), i.e. K is expressed in terms of the function k of two variables $x \in \mathbb{R}^d$, $y \in \mathbb{R}^d$ through the formula

$$K(x,A) = \int_{y \in A} k(x,y)\, dy.$$

In that case, if $\mu^{(n)}$ (i.e. X_n) has a density $\pi^{(n)}(x)$, then $\mu^{(n+1)}$ has a density (see (12))

$$\pi^{(n+1)}(y) = \int_{\mathbb{R}^d} \pi^{(n)}(x)\, dx k(x,y).$$

The density of a stationary distribution satisfies (see (13))

$$\pi(y) = \int_{\mathbb{R}^d} \pi(x)\, dx k(x,y).$$

and the detailed balance condition (14) becomes

$$\pi(x)k(x,y) = \pi(y)k(y,x).$$

5.2 Accept/Reject with Continuous Targets

If the target has a density π, the Metropolis acceptance probability for the discrete case given in (10) has to be replaced by

$$a = 1 \wedge \frac{\pi(x^*_{n+1})}{\pi(x_n)}, \tag{15}$$

where x_n is the value of X_n (current location of the chain) and x^*_{n+1} is the proposal for the next location. This formula requires that the proposal be based on a symmetric kernel (in the case with densities the density of the proposal kernel must satisfy $k^*(x,y) \equiv k^*(y,x)$). The Hastings formula (11) may be similarly adapted.

5.3 The Random Walk Metropolis Algorithm

As mentioned in the discrete case, a simple way of generating proposals is to use a random walk format: $Y_{n+1} = Y_n + Z_n$ where now the Z_n are independent identically distributed continuous random variables in $\mathbb{R}^d$. A common choice is to take each Z_n to be a *normal (Gaussian)* d-variate distribution $N(m,C)$ with density given by:

$$\frac{1}{(2\pi)^{d/2}\det(C)^{1/2}}\exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right). \tag{16}$$

Here $m \in \mathbb{R}^d$ is the expectation $\mathbb{E}(Z_n)$ and C is the $d \times d$ symmetric positive definite matrix of the covariances of the d (scalar) components $Z_{n,i}$ of Z_n, i.e.

$$c_{ij} = \mathbb{E}((Z_{n,i} - m_i)(Z_{n,j} - m_j)).$$

Of course in the scalar ($d = 1$) case, (16) becomes

$$\frac{1}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-m)^2\right), \tag{17}$$

where σ^2 is the variance.

For $m = 0$ the normal distribution (16) is symmetric and therefore the proposal satisfies the symmetry condition required to apply the Metropolis accept/reject formula (15). The overall procedure is then called a Metropolis Random Walk (RW) algorithm. In the absence of other information, it is reasonable to use in (16) a scalar covariance matrix $C = h^2 I_d$ (so that the scalar components $Z_{n,i}$ of the random vector Z_n have a common variance h^2 and are uncorrelated). Then the RW proposal is

$$X^*_{n+1} = X_n + hZ_n, \qquad Z_n \sim N(0, I_d). \tag{18}$$

Let us present an example. Assume that the (univariate) target probability density is[15]

$$\propto \exp(-\beta V(x)), \qquad V(x) = x^4. \tag{19}$$

Anticipating here some elements of the discussion in Section 8.3, we mention that (19) would arise in statistical physics as the Boltzmann density for a system consisting of a single particle in a potential well with potential $V(x)$ interacting thermally with the environment at absolute temperature $\propto 1/\beta$ (more precisely $\beta = 1/(k_B T_a)$ where k_B is Boltzmann's constant and T_a the absolute temperature). If β is close to ∞ (low temperature) the particle will be at the location $x = 0$ of minimum potential energy. As the temperature increases, the particle is hit e.g. by moving molecules in the environment, and it may leave the minimum $x = 0$. In an ensemble of such systems the value of x will be distributed as in (19).

We have applied to the target (19), the RW algorithm (18). With $\beta = 1$, $h = 1$ and $N = 1{,}000{,}000$ steps, we obtained the histogram in Fig. 6.

Of course the N correlated samples of the target generated by the algorithm contain less information than N independent samples would afford and, therefore, high correlation impairs the usefulness of the samples delivered by the algorithm. In this connection, the choice of the standard deviation h in (18) has a marked effect on the performance of RW. A lower value of h leads to fewer rejections, but the progress of

[15] The symbol $\propto$ means proportional to. To obtain a probability density it is necessary to divide $\exp(-\beta V(x))$ by the normalizing constant $\int_{\mathbb{R}} V(x)\,dx$. As pointed out before the Metropolis algorithm does not require the knowledge of the normalizing constant.

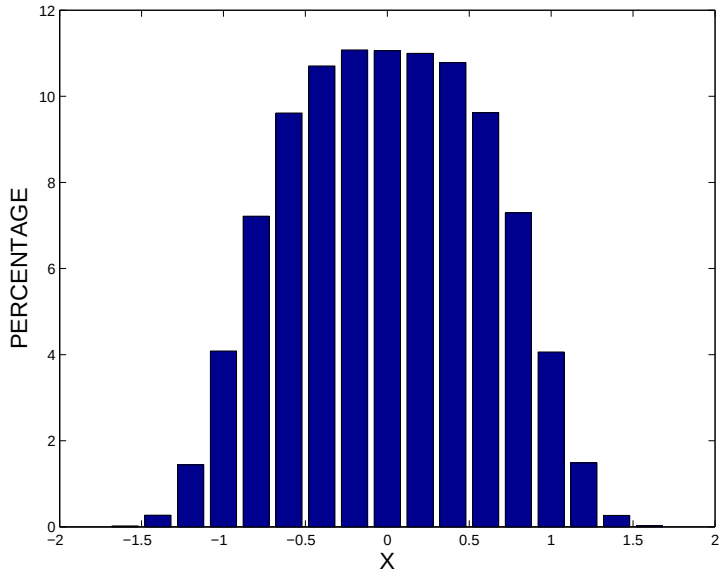

Fig. 6 Histogram of the target density $\propto \exp(-x^4)$ obtained by the RW algorithm with $h = 1$

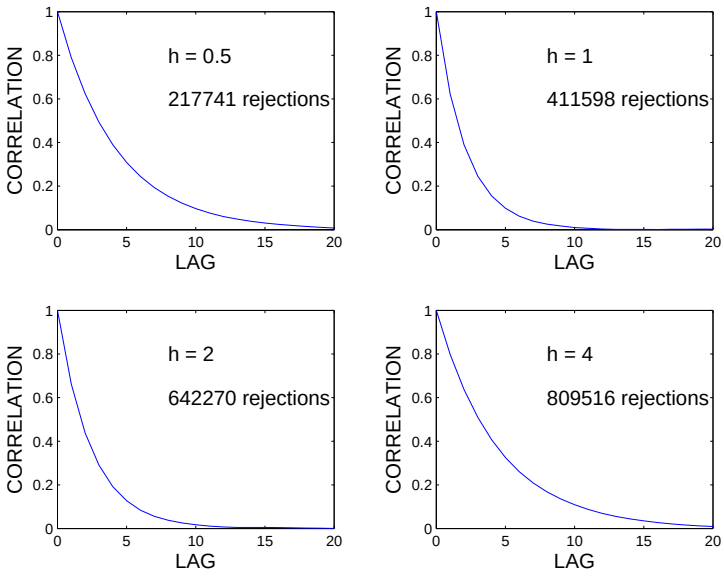

Fig. 7 RW: decay of the empirical correlation coefficient ρ_ν between X_n and $X_{n+\nu}$ as a function of the lag ν

the chain is then slow (i.e. the locations x_{n+1} and x_n at consecutive steps are close to one another). Therefore the correlation between the random variables X_n and X_{n+1} is then high. Large values of h lead to more frequent rejections. Since at a rejected step the chain does not move, $x_{n+1} = x_n$, this also causes an increase of the correlation between X_n and X_{n+1}.

The empirical *auto-covariance* with lag ν of the samples $x_0, \ldots, x_N$ given by the algorithm is

$$\gamma_\nu = \frac{1}{N+1} \sum_{i=0}^{N-\nu} \big(x_i - \widehat{m}\big)\big(x_{i+\nu} - \widehat{m}\big), \qquad \widehat{m} = \frac{1}{N+1} \sum_{i=0}^{N} x_i,$$

and accordingly $\rho_\nu = \gamma_\nu/\gamma_0$ represents the empirical *auto-correlation coefficient*. We would then like that the value ρ_ν approaches zero as quickly as possible as ν increases. Fig. 7 illustrates the behavior of ρ_ν as h varies. Note that the number of rejections increases with h and that $h = 1, 2$ are the best choices among those considered.[16]

6 Stochastic Differential Equations

In the RW algorithm the choice of proposals is completely independent of the target distribution. It is plausible that, by incorporating into the proposals some knowledge of the target, MCMC algorithms may take large steps from the current position without drastically reducing the chance of having the proposal accepted. Stochastic differential equations (SDEs) provide a means to improve on random walk proposals.

The rigorous study of continuous-time stochastic processes in general and of SDEs in particular is rather demanding mathematically. Here, in the spirit of [18], we present an algorithmic introduction, focused on how to simulate such processes in the computer. This kind of simulation provides much insight and is a very useful first step for those wishing to study these issues.

6.1 The Brownian Motion

It is well known that the Brownian motion of pollen grains in water is named after the Scottish botanist R. Brown who described it 1827. For three quarters of a century the causes of the motion remained unclear, until in 1905 A. Einstein offered a complete explanation in terms of shocks provided by the molecules of water, thus furnishing the definitive proof of the molecular nature of matter. The mathematical Brownian motion (also called the Wiener process or Wiener-Bachelier process) was

[16] For further details of the statistical analysis of the sequence of samples x_i the reader is referred to [13].

first studied by Bachelier in 1900 and then by Wiener in the 1920's. The standard or normalized, scalar Brownian motion is a real-value stochastic process B_t ($t \in [0,\infty)$) with the following characteristic features, [12], Chapter 3, [5], Section 37:

1. $B_0 = 0$.
2. It has independent increments (i.e. if $0 \leq s_1 < t_1 < s_2 < t_2$, the random variables $B_{t_1} - B_{s_1}$ and $B_{t_2} - B_{s_2}$ are independent).
3. For $t > s$, $B_t - B_s \sim N(0, t-s)$,[17] $0 \leq s < t$ (see (16)).
4. It has continuous paths $t \mapsto B_t$.

A d-dimensional Wiener process takes values in $\mathbb{R}^d$ and its components are independent one-dimensional Wiener processes. The mathematical construction of B may be seen e.g. in [12], Chapter 3, [5], Section 37.[18]

After discretizing the variable t on a grid $t = 0, \Delta t, 2\Delta t, \ldots$, the d-dimensional Wiener process may be simulated [18] by the recursion

$$B_{n+1} = B_n + \sqrt{\Delta t} Z_n,$$

where B_n is the approximation corresponding to $t_n = n\Delta t$ and the Z_n are independent variables with distribution $N(0, I_d)$ (see (16)). Note that the simulated (discrete time) process $\{B_n\}$ has then independent Gaussian increments with the right expectation $\mathbb{E}(B_n) = 0$ and variance $\mathrm{Var}(B_n) = t_n$.

Fig. 8 depicts two simulated trajectories for $0 \leq t \leq 1$, $\Delta t = 0.0001$. Since over a time interval of small length Δt the increment $B_{n+1} - B_n$ has the relatively large standard deviation $\sqrt{\Delta t}$, simulated paths have a rugged look. In fact, before discretization, the Wiener paths are almost surely nowhere differentiable, [5] Theorem 37.3, [12] Chapter 3, Theorem 2.2.[19]

6.2 The Euler-Maruyama Method

A stochastic differential equation has the form ([24] Chapter 2, [12] Chapter 5)

$$dX_t = f(X_t, t)dt + \sigma(X_t, t)dB_t, \tag{20}$$

where f takes values in $\mathbb{R}^d$, σ takes values in the $d \times d'$ real matrices and B_t is a d'-dimensional Wiener process. The first term in the right-hand side provides a deterministic *drift;* the second random *noise or diffusion*. The expression dB_t does not make sense, since, as pointed out above, the paths $t \mapsto B_t$ are non-differentiable. In fact the differential equation is shorthand for the integral equation

[17] $\sim$ means 'has a distribution.'

[18] These references also show that 4. is essentially a consequence of 1., 2. and 3.

[19] The trajectories of the Wiener process are in fact complex objects. For instance, with probability 1, the set $Z(\omega)$ of values t for which a trajectory $B_t(\omega)$ vanishes is closed, unbounded, without isolated points and of Lebesgue measure 0, [5], Theorem 37.4.

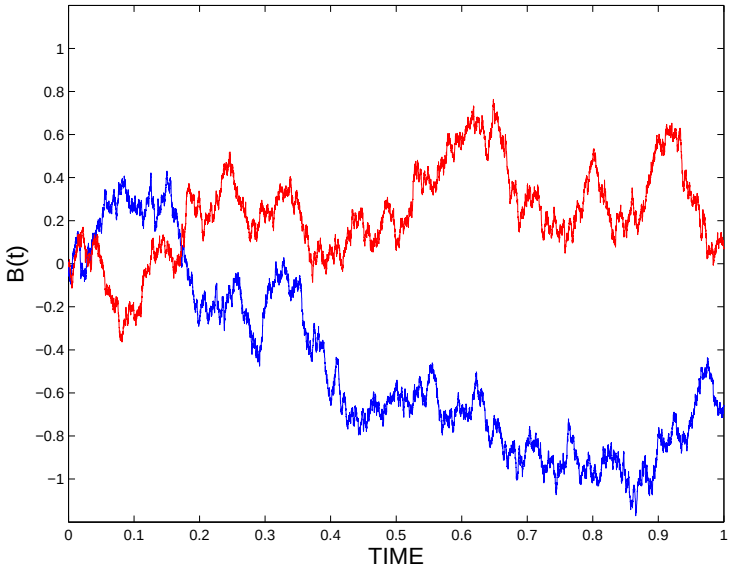

Fig. 8 Two simulated trajectories of the standard Brownian motion

$$X_t = X_0 + \int_0^t f(X_s,s)ds + \int_0^t \sigma(X_s,s)dB_s,$$

where the last term is an Ito integral ([12] Chapter 4, [24] Chapter 1). Simulations may be carried out by the Euler-Maruyama discretization [18]:

$$X_{n+1} = X_n + \Delta t f(X_n, n\Delta t) + \sqrt{\Delta t}\sigma(X_n, n\Delta t)Z_n,$$

where the Z_n are independent $\sim N(0, I_{d'})$.

As a simple example consider the scalar problem

$$dX_t = X_t dt + dB_t, \quad t > 0, \qquad X_0 = 1 \tag{21}$$

(the initial condition here is deterministic, but cases where X_0 is a random variable often occur). Without noise, the solution would of course be $X_t = \exp(t)$. The Euler-Maruyama discretization is ($Z_n \sim N(0,1)$):

$$X_{n+1} = X_n + \Delta t X_n + \sqrt{\Delta t} Z_n, \quad n = 0, 1, \ldots, \qquad X_0 = 1.$$

Ten trajectories of the simulated solution X_t, $0 \le t \le 1$, may be seen in Fig. 9 ($\Delta t = 0.0001$). Clearly the paths exhibit an upward drift, corresponding to the tern $X_t dt$ in (21), while at the same time showing a diffusion whose variance increases with t. In order to visualize the drift better, we have simulated $100,000$ trajectories in $0 \le t \le 1$ with step-size $\Delta t = 0.001$, recorded the value of $X_{t=1}$ for each trajectory and

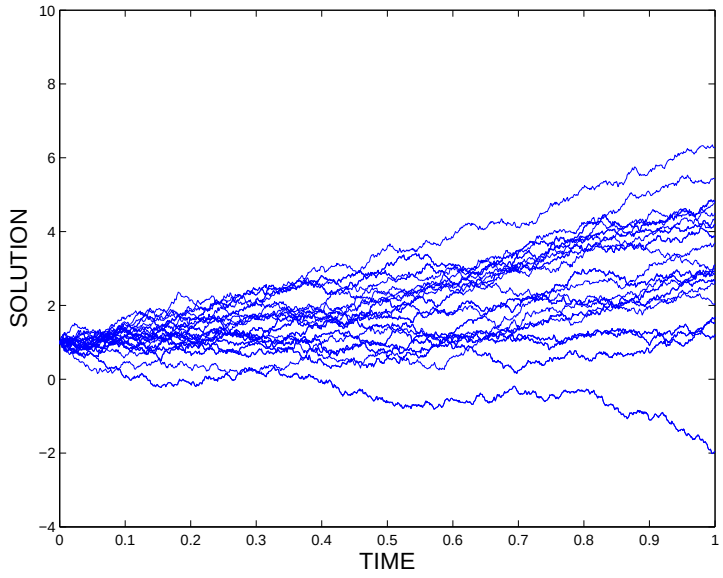

Fig. 9 Ten trajectories of the stochastic differential equation (21)

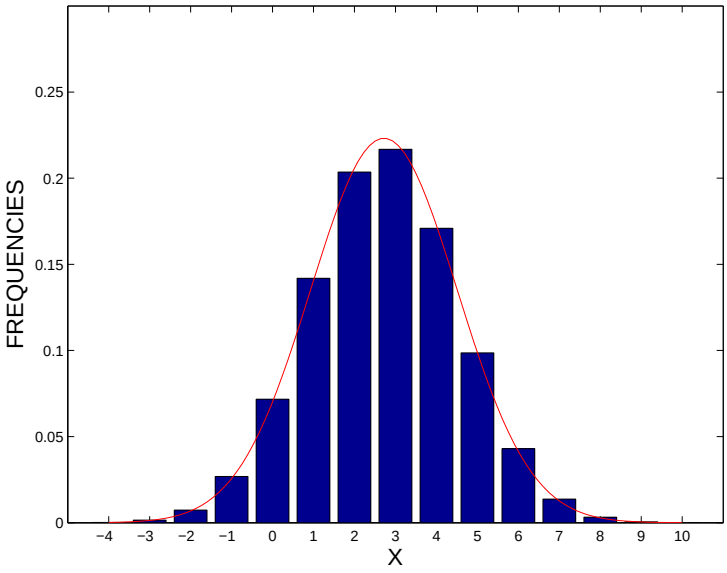

Fig. 10 Histogram of 100,000 samples of $X_{t=1}$, where X_t is the solution of the stochastic differential equation (21). The line is the Gaussian density with mean e and variance $(e^2-1)/2$

produced an histogram by distributing those $100,000$ values into 15 bins centered at $-4,-3,\ldots,10$. The result may be seen in Fig. 10. A Gaussian density with mean e and variance $(e^2-1)/2$ provides an excellent fit to the distribution of $X_{t=1}$.

6.3 The Fokker-Planck Equation

How did we find the probability distribution of the solution X_t at $t=1$? The densities $\pi(x,t)$, $x\in\mathbb{R}^d$, of the solution X_t of the SDE (20) obey the *Fokker-Planck* equation [23] Section 2.2.1[20]

$$\partial_t \pi(x,t) + \sum_{i=1}^{d} \partial_i \Big(f^i \, \pi(x,t) \Big) = \frac{1}{2} \sum_{i,j=1}^{d} \partial_i \partial_j \Big(a^{i,j} \, \pi(x,t) \Big), \tag{22}$$

where $a=\sigma\sigma^T$, superscripts denote components and f and a are of course evaluated at (x,t). Let us see some examples.

6.3.1 Fokker-Planck Equation: No Drift

Consider first the scalar equation without drift: $dX_t = dB_t$. When $X_0=0$ the solution is of course B_t and may be seen as describing the abscissa of a particle that moves due to random shocks from the environment. The Fokker-Planck equation (22) is the familiar heat equation

$$\partial_t \pi(x,t) = \frac{1}{2} \partial_{xx} \pi(x,t);$$

this governs the diffusion of the trajectories of the particle corresponding to different realizations of the process or, in the language of ensembles, the diffusion of an ensemble of particles, initially located at the origin, that evolve randomly independently from one another.

If the initial condition for the partial differential equation is a unit mass located at $x=0$ (the initial location of the particle is 0 independently of the chance), then the solution, by definition, is the *fundamental solution of the heat equation,* which has the familiar expression:

$$\pi(x,t) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2} \frac{x^2}{t} \right).$$

Comparing with (17), we conclude that $X_t \sim N(0,t)$; this of course matches the fact that $B_t \sim N(0,t)$ as we know.

[20] The terminology Fokker-Panck is used in physics; in probability the equation is known as Kolmogorov's forward equation, see e.g. [10] Chapter X, Section 5

6.3.2 Fokker-Planck Equation: No Diffusion

Assume now that the SDE is the standard ordinary differential equation (ODE) $dX_t = f(X_t,t)dt$ so that the dynamics are deterministic. The Fokker-Planck equation (22) reads

$$\partial_t \pi + \nabla \cdot (\pi f) = 0,$$

where we recognize the familiar Liouville equation for the transport of densities by the ODE (see e.g. [21], Section 10.1). The trajectories of the ODE are characteristic curves of this first-order linear partial differential equation.

6.3.3 Fokker-Planck Equation: A Linear Example

For the problem (21), the Fokker-Planck equation (22) is given by

$$\partial_t \pi + \partial_x (x\pi) = \frac{1}{2}\partial_{xx}\pi.$$

The solution with initial data given by a unit mass at $x = 1$ is found to be:

$$\frac{1}{\sqrt{2\pi}\sigma(t)} \exp\left(-\frac{1}{2}\frac{(x-m(t))^2}{\sigma^2(t)}\right),$$

with

$$m(t) = \exp(t), \qquad \sigma^2(t) = \frac{\exp(2t)-1}{2}.$$

Comparison with (17) shows that the solution has, at each t, the Gaussian distribution with variance $\sigma^2(t)$ and expectation $m(t)$ (this average $\exp(t)$ coincides with the solution when the noise is turned off so that the equation becomes $dX_t = X_t dt$).

7 Metropolis Adjusted Langevin Algorithm

The Metropolis adjusted Langevin algorithm (MALA) [31] is an instance of a MCMC where proposals are based on an SDE for which the target π is an invariant density, cf. [30] Section 7.8.5. Without loss of generality (densities are positive) the target density is written as $\pi \propto \exp(\mathscr{L})$. Then the proposal is (cf. (18))

$$X^*_{n+1} = X_n + \frac{h^2}{2}\nabla \mathscr{L}(X_n) + hZ_n, \qquad Z_n \sim N(0, I_d).$$

The middle term in the right-hand side (absent in the RW proposal) provides an increment in the direction of steepest ascent in $\mathscr{L}$, thus biasing the exploration of the state space towards high-probability areas. Since the proposal kernel is not symmetric, the Hastings accept/reject mechanism must be used.

More generally, one may use a *preconditioned* version of the proposal

$$X^*_{n+1} = X_n + \frac{h^2}{2} M^{-1} \nabla \mathscr{L}(X_n) + h\sqrt{M^{-1}} Z_n,$$

with M a symmetric positive definite $d \times d$ constant matrix. The idea is that M should be taken 'large' in those directions in state space where smaller increments are desirable because the target varies more quickly.

The recipe for the proposal is an Euler-Maruyama step with step-length $\Delta t = \sqrt{h}$ for the SDE

$$dX_t = \frac{1}{2} M^{-1} \nabla \mathscr{L}(X_t) dt + \sqrt{M^{-1}} dB_t$$

whose Fokker-Planck equation has the target $\propto \exp(\mathscr{L}(x))$ as a stationary (time-independent) solution. This implies that, at stationarity in the chain, the proposals will be distributed (except for the Euler discretization error) according to the target: therefore high acceptance rates will be attained. In practice it is of course impossible to start the chain from the stationary distribution, but chains are likely to be asymptotically stationary (Section 3.6) and high acceptance rates may be expected in that case.

The paper [32] proves that if the target consists of d independent copies of the same distribution then the RW algorithm requires $h \propto 1/d$ to have $\mathscr{O}(1)$ acceptance probabilities as $d \to \infty$. MALA improves on that because, as shown in [33], it may operate with larger values of h, namely $h \propto (1/d)^{1/3}$. These papers also show that these algorithms perform best when the acceptance probability is approximately 0.234 for the RW case and 0.574 for MALA. These results have recently been extended [25], [29] to situations where the target is not product of equal copies.

As an example of the use of MALA, consider target density

$$\propto \exp(-(1/2)k(r-1)^2), \qquad r = |x|, \qquad x \in \mathbb{R}^d.$$

This is the Boltzmann density (Section 8.3) for the motion of a spring in d dimensions. Here we take $d = 100$, $k = 100$; the the probability is concentrated in the neighborhood of the unit sphere $|r| = 1$, i.e. essentially on a manifold of dimension 99. We have applied the RW and MALA algorithms. Figs. 11 (RW) and 12 (MALA) show the projections of the draws x onto the two-dimensional plane (x_1, x_2) (the corresponding marginal distribution is concentrated in the neighborhood of the origin in the (x_1, x_2)-plane) and the correlation in the variable x_1. Clearly MALA is able to take larger steps allowing for a faster exploration of the distribution.

8 Hamiltonian Dynamics

We now turn our attention to Hamiltonian systems, a topic that is essential to formulate the Hybrid Monte Carlo method to be discussed in the next section.

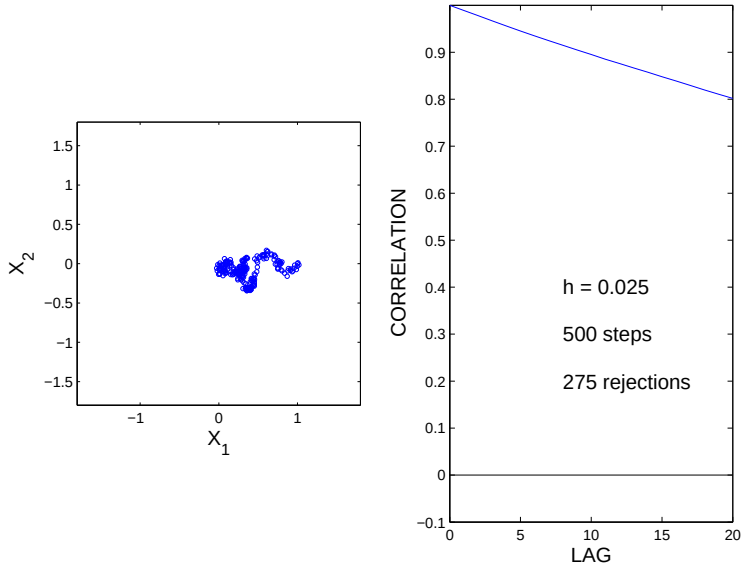

Fig. 11 RW results for a stiff spring in $\mathbb{R}^{100}$. Samples of the coordinates x_1, x_2 and autocorrelation in x_1

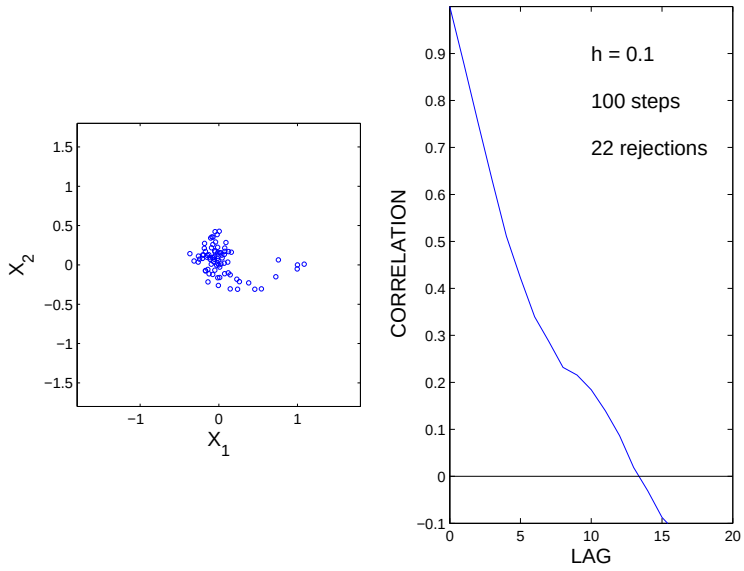

Fig. 12 MALA results for a stiff spring in $\mathbb{R}^{100}$. Samples of the coordinates x_1, x_2 and autocorrelation in x_1

8.1 An Example: Systems of Point Masses

The study of the motion of a conservative system of point masses in three-dimensional space is of much interest in many branches of science. Examples range from molecular dynamics, where the particles are molecules or atoms, to astrophysics, where one deals with stars or galaxies. If ν is the number of particles, m_i the mass of the i-th particle, and $\mathbf{r}_i \in \mathbb{R}^3$ its radius vector, Newton's second law reads:

$$m_i \frac{d^2}{dt^2}\mathbf{r}_i = -\nabla_i V(\mathbf{r}_1, \ldots, \mathbf{r}_\nu), \quad i = 1, \ldots, \nu,$$

where the scalar V is the *potential* and $-\nabla_i V$ is the net force on the i-th particle (∇_i means gradient with respect to $\mathbf{r}_i$). This is of course a system of 3ν *second-order* scalar differential equations for the 3ν cartesian components $r_{i,j}$ of the $\mathbf{r}_i$, $j = 1, 2, 3$. After introducing the *momenta*

$$\mathbf{p}_i = m_i \frac{d}{dt}\mathbf{r}_i, \quad i = 1, \ldots, \nu, \tag{23}$$

the equations may be rewritten in *first-order* form:

$$\frac{d}{dt}\mathbf{p}_i = -\nabla_i V(\mathbf{r}_1, \ldots, \mathbf{r}_\nu), \quad i = 1, \ldots, \nu.$$

These 3ν scalar equations, together with the 3ν scalar equations in (23) provide a system of $D = 2 \cdot 3 \cdot \nu$ first-order scalar differential equations for the D cartesian components of the vectors $\mathbf{r}_i$ and $\mathbf{p}_i$, $i = 1, \ldots, \nu$. With the introduction of the *Hamiltonian* function

$$H = \sum_i \frac{1}{2m_i}\mathbf{p}_i^2 + V(\mathbf{r}_1, \ldots, \mathbf{r}_\nu) \tag{24}$$

the first-order system takes the very symmetric *canonical* form:

$$\frac{d}{dt}p_{i,j} = -\frac{\partial H}{\partial r_{i,j}}, \quad \frac{d}{dt}r_{i,j} = +\frac{\partial H}{\partial p_{i,j}}, \quad i = 1, \ldots, \nu, \quad j = 1, 2, 3. \tag{25}$$

Note that H represents the total mechanical *energy* in the system, composed of a *kinetic* part

$$\sum_i \frac{1}{2m_i}\mathbf{p}_i^2 = \sum_i \frac{1}{2} m_i \left(\frac{d}{dt}\mathbf{r}_i\right)^2$$

and a *potential* part V.

The use of the Hamiltonian format in lieu of Newton's equations is essential in statistical mechanics and quantum mechanics.

8.2 Hamiltonian Systems

In the *phase space* $\mathbb{R}^D$, $D = 2d$, of the points (p,x),

$$p = (p_1, \ldots, p_d) \in \mathbb{R}^d, \qquad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

to each smooth real-valued function $H = H(p,x)$ (the Hamiltonian) there corresponds a first-order differential system of D *canonical* Hamiltonian equations (cf. (25)):

$$\frac{d}{dt}p_j = -\frac{\partial H}{\partial x_j}, \qquad \frac{d}{dt}x_j = +\frac{\partial H}{\partial p_j}, \qquad j = 1, \ldots, d. \tag{26}$$

In mechanics, as it was the case in (25), the variables $x \in \mathbb{R}^d$ describe the *configuration* of the system, the variables p are the momenta conjugate to x [2] and d is the number of *degrees of freedom*.

Canonical Hamiltonian systems appear very frequently in science; virtually all situations where dissipation is absent or may be neglected may be brought into Hamiltonian form. At the same time, Hamiltonian systems possess properties not frequently found in 'general' systems. Before discussing such special properties, it is convenient to introduce the *flow* $\{\Phi_t\}_{t\in\mathbb{R}}$ of the system (26). For each fixed (but arbitrary) t (see [35], Section 2.1), Φ_t is a map in phase space, $\Phi_t : \mathbb{R}^D \to \mathbb{R}^D$, defined as follows: for each point (p_0,x_0), $\Phi_t(p_0,x_0)$ is the value at time t of the solution $(p(t),x(t))$ of the canonical equations (26) with value (p_0,x_0) at time 0. The simplest example has $d = 1$ and $H = (1/2)(p^2 + x^2)$, the canonical system is $(d/dt)p = -x$, $(d/dt)x = p$ (the harmonic oscillator). The solution with initial value (p_0,x_0) is

$$p(t) = p_0 \cos t - x_0 \sin t, \qquad x(t) = p_0 \sin t + x_0 \cos t,$$

and therefore Φ_t is the rotation in the plane that moves the point (p_0,x_0) to the point

$$\Phi_t(p_0,x_0) = \big(p_0 \cos t - x_0 \sin t,\, p_0 \sin t + x_0 \cos t\big);$$

$\{\Phi_t\}_{t\in\mathbb{R}}$ is the one-parameter family of rotations in the plane. Note that, in general, $\Phi_t(p_0,x_0)$ means:

- If t is varied while keeping (p_0,x_0) fixed: the solution of (26) with initial condition (p_0,x_0).
- If t is fixed and (p_0,x_0) regarded as a variable: a transformation Φ_t in phase space.
- If t is regarded as a parameter: a one-parameter family $\{\Phi_t\}_{t\in\mathbb{R}}$ of transformations in phase-space. This family is a *group*: $\Phi_t \circ \Phi_s = \Phi_{t+s}$, $\Phi_{-t} = \Phi_t^{-1}$.

We now describe the properties of Hamiltonian systems that we shall require when formulating and analyzing the Hybrid Monte Carlo method.

8.2.1 Properties of Hamiltonian Systems: Conservation of Energy

The function H is a *conserved quantity* or *first integral* of (26). In fact, along solutions:

$$\begin{aligned}\frac{d}{dt}H(p(t),x(t)) &= \sum_j \left(\frac{\partial H}{\partial p_j}\frac{d}{dt}p_j + \frac{\partial H}{\partial x_j}\frac{d}{dt}x_j\right)\\ &= \sum_j \left(-\frac{\partial H}{\partial p_j}\frac{\partial H}{\partial x_j} + \frac{\partial H}{\partial x_j}\frac{\partial H}{\partial p_j}\right) = 0,\end{aligned}$$

and therefore

$$H(p(t),x(t)) = H(p(0),x(0).$$

In terms of the flow, this property simply reads $H \circ \Phi_t = H$ for each t.

For the example (25) with ν point masses, we pointed out that H measures the total energy; therefore the conservation of H corresponds to *conservation of energy*. This is also the case for most Hamiltonian problems.

8.2.2 Properties of Hamiltonian Systems: Conservation of Volume

For each t, Φ_t is a *volume preserving* transformation in phase space ([35], Section 2.6): for each (Borel) subset $A \subset \mathbb{R}^D$,

$$\mathrm{Vol}\big(\Phi_t(A)\big) = \mathrm{Vol}(A).$$

For the simple example of the harmonic oscillator, this corresponds to the obvious fact that the area of a planar set A does not change when the set is rotated. In general, conservation of volume is a direct consequence of Liouville's theorem: the solution flow of a differential system $\dot{z} = G(z)$ is volume preserving if and only if the corresponding vector field G is divergence-free ([2] Section 16). Indeed for a canonical system the divergence is

$$\sum_{j=1}^{d}\left(\frac{\partial}{\partial p_j}\left(-\frac{\partial H}{\partial x_j}\right) + \frac{\partial}{\partial x_j}\frac{\partial H}{\partial p_j}\right) = 0.$$

In terms of the flow, conservation of volume simply reads: $\det(\Phi_t') \equiv 1$, for each t where Φ_t' denotes the Jacobian of Φ_t.

8.2.3 Properties of Hamiltonian Systems: Reversibility

Consider now the momentum-flip symmetry S in phase-space defined by

$$S(p,x) = (-p,x)$$

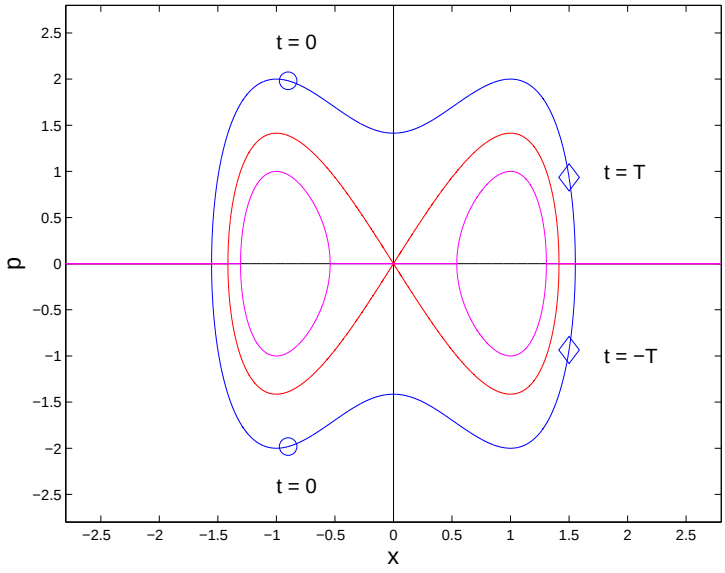

Fig. 13 Reversibility: $\Phi_t \circ S = S \circ \Phi_{-t}$. Begin from the lower circle, flip the momentum to get the upper circle and use the solution flow to reach the upper diamond after T units of time. The final result is the same as one would get by first evolving $-T$ units of time to reach the lower diamond and then flipping the momentum

and assume that $H \circ S = H$, i.e. the Hamiltonian is an even function of the momenta as in (24) and many other mechanical systems. If $(p(t),x(t))$ is a solution of the canonical equations (26), so is $(\widehat{p}(t),\widehat{x}(t)) := (-p(-t),x(-t))$. The proof is simple:

$$\frac{d}{dt}\widehat{p}_i(t) = \frac{d}{dt}p_i(-t) = -\frac{\partial H}{\partial x_i}(p(-t),x(-t)) = -\frac{\partial H}{\partial x_i}(\widehat{p}(t),\widehat{x}(t))$$

and similarly for x_i. Since $(\widehat{p}(0),\widehat{x}(0)) = S(p,x)$, this fact, called *reversibility* of the flow, may be compactly written as

$$\Phi_t \circ S = S \circ \Phi_{-t};$$

see Fig. 13 that portraits the two-dimensional phase space of the Hamiltonian function

$$H(p,x) = \frac{1}{2}p^2 + V(x), \qquad V(x) = (x^2-1)^2. \tag{27}$$

The significance of reversibility is well known: if a movie is made of a motion of a reversible system and projected backwards what we see is also a possible (forwards) motion of the system. In Fig. 13, if the forwards movie shows the sequence of configurations $x(t)$ from $t = 0$ to $t = T$ (top circle to top diamond), when projected backwards will display in reversed order the same configurations and that

sequence of configurations corresponds to the solution that at time $t = 0$ starts at the lower diamond and reaches the lower circle at $t = T$.

8.3 The Canonical Density

Consider a mechanical system whose time-evolution, when isolated from rest of the universe, is governed by the canonical equations associated with a Hamiltonian $H(p,x)$ (for instance our earlier system of ν point masses). As discussed above, the value of $H(p(t),x(t))$ (the energy) along the evolution remains constant. Assume now that this system is not isolated but interacts with an environment at constant temperature (for instance our point masses collide with water molecules of a heat bath that surrounds them). There will be exchanges of energy between the system and the environment, the value of H will nort remain constant and the canonical equations (26) will not describe the time evolution. The energy exchanges with the environment must be modeled as random and therefore, for each fixed t, $(p(t),x(t))$ are random variables that need to be described statistically. Once thermal equilibrium with the environment has been reached, the stochastic process $(p(t),x(t))$ possesses a stationary probability distribution: this is the *Maxwell-Boltzmann* distribution with density

$$\propto \exp(-\beta H(p,x)), \qquad \beta = \frac{1}{k_B T_a} \tag{28}$$

(here T_a is the absolute temperature of the environment and k_B the Boltzmann constant). The corresponding ensemble is called the *canonical ensemble* ([36] Section 12.5.3, [21] Section 10.2). Thus, at any temperature, a canonical ensemble contains few systems at locations of high energy. However, as the temperature T_a increases locations of high energy become more likely.

8.3.1 The Maxwell-Boltzmann Distribution

As an example consider again the system of point masses with Hamiltonian (24). The canonical density (28) is

$$\propto \exp\left(-\beta\left(\sum_{i=1}^{\nu}\frac{1}{2m_i}\mathbf{p}_i^2 + V(\mathbf{r}_1,\ldots,\mathbf{r}_\nu)\right)\right),$$

and since the exponential may be rewritten as a product, the $\nu+1$ random vectors $\mathbf{p_1} \in \mathbb{R}^3,\ldots,\mathbf{p}_\nu \in \mathbb{R}^3$, $(\mathbf{r}_1,\ldots,\mathbf{r}_\nu) \in \mathbb{R}^{3\nu}$ are mutually independent. The density of $\mathbf{p}_i$ is then ([11] Section 40.4), as first established by Maxwell in 1859,

$$\propto \exp(-\frac{\beta}{2m_i}\mathbf{p}_i^2),$$

and comparison with (16) shows that the distribution of $\mathbf{p}_i$ is Gaussian with zero mean and covariance matrix $(m_i/\beta)I_3$. In particular there is no correlation among the three cartesian components $p_{i,j}$ of $\mathbf{p}_i$ and each of these components has variance $m_i/\beta = m_i k_B T_a$. It follows that the kinetic energy $p_{i,j}^2/(2m_i)$ of the i-th mass along the j-axis has an ensemble average $(1/2)k_B T_a$; in other words the absolute temperature coincides, up to a normalizing factor, with the kinetic energy in any of the 3ν degrees of freedom of the system (in fact this is the *definition* of absolute temperature [11] Section 39.4).

The configuration, specified by $(\mathbf{r}_1,\ldots,\mathbf{r}_\nu) \in \mathbb{R}^{3\nu}$, is, as noted above, independent of the momenta, and possesses the density

$$\propto \exp\left(-\beta V(\mathbf{r}_1,\ldots,\mathbf{r}_\nu)\right).$$

In statistical mechanics this is called the *Boltzmann* density for the potential V ([11] Section 40.2).

8.3.2 Preservation of the Canonical Density by the Hamiltonian Flow

We shall need later the following result:

Theorem 7. *For each fixed t, the canonical density (28) is preserved by the flow Φ_t of the Hamiltonian system (26):*

$$\int_{\Phi_t(A)} \exp(-\beta H(p,x))\,dp\,dx = \int_A \exp(-\beta H(p,x))\,dp\,dx,$$

for each (Borel) subset A of the phase space $\mathbb{R}^D$.

Proof. Change variables $(p,x) = \Phi_t(\tilde{p},\tilde{x})$ in the first integral; $H(\Phi_t(\tilde{p},\tilde{x})) = H(\tilde{x},\tilde{p})$ by conservation of energy and the required Jacobian determinant is unity by conservation of volume. □

In an image due to Gibbs, one places at $t = 0$ points in the phase space $\mathbb{R}^D$ of H in such a way that they are distributed with probability density $\propto \exp(-\beta H)$. Each point represents a system in the canonical ensemble. As t varies each point will move in the phase space following (26); the theorem implies that the density at any point (p,x) will remain constant.

8.4 Numerical Methods for Hamiltonian Problems

The analytic integration of Hamilton's canonical equations (26) is usually impossible and one has to resort to numerical integrators. In the last twenty five years it has become clear that when integrating Hamiltonian problems it is essential in many applications to use numerical methods that possess conservation properties similar to

those shared by Hamiltonian systems, like reversibility, conservation of volume, etc. The construction and analysis of such numerical methods is part of the field of *geometric integration,* a term coined in [34]. An introductory early monograph is [35] and a more recent expositions are given in [16] and [22]. Here we limit ourselves to the material required later to describe the Hybrid Monte Carlo method.

Each one-step *numerical method* to integrate (26) is specified by a map $\psi_{\Delta t} : \mathbb{R}^D \to \mathbb{R}^D$, where Δt represents the step-length. The approximation (p^{m+1}, x^{m+1}) to the true solution value $(p((m+1)\Delta t), x((m+1)\Delta t))$ is obtained recursively by

$$(p^{m+1}, x^{m+1}) = \psi_{\Delta t}(p^m, x^m).$$

In this way the approximate solution at time T (for simplicity we assume that $T/\Delta t$ is an integer) is obtained by applying $T/\Delta t$ times the mapping $\psi_{\Delta t}$, or, in other words, the true T-flow $\Phi_T = \Phi_{\Delta t}^{T/\Delta t}$ is approximated by $\Psi_T := \psi_{\Delta t}^{T/\Delta t}$, the composition of $\psi_{\Delta t}$ $T/\Delta t$ times with itself.

It turns out that it is impossible to construct a general integrator $\psi_{\Delta t}$ that exactly preserves energy and volume ([35] Section 10.3.2). Faced with this impossibility, it is advisable to drop the requirement of exact conservation of energy and demand exact conservation of volume.[21] The best-known volume preserving, reversible algorithm to integrate Hamiltonian systems is the Verlet/Stoermer/leapfrog algorithm applicable when the Hamiltonian has the special (but common) form (cf. (24))

$$H = \frac{1}{2} p^T M^{-1} p + V(x);$$

M a constant positive definite symmetric matrix —the so-called mass matrix—. For our purposes we may think of this method as a *splitting* (fractional step) algorithm, see [35], Section 12.4. The Hamiltonian H is written as a sum $H = H^{(1)} + H^{(2)}$ of potential and kinetic parts with

$$H^{(1)}(p,x) = V(x), \quad H^{(2)}(p,x) = \frac{1}{2} p^T M^{-1} p.$$

For the Hamiltonian $H^{(1)}$, the equations of motion are $(d/dt)p = -(\partial/\partial x)V(x)$, $(d/dt)x = 0$, with solutions

$$p(t) = p_0 - t\frac{\partial}{\partial x} V(x_0), \qquad x(t) = x_0.$$

For the Hamiltonian $H^{(2)}$, the equations of motion are $(d/dt)p = 0$, $(d/dt)x = M^{-1}p$ leading to

$$p(t) = p_0, \qquad x(t) = x_0 + tM^{-1}p_0.$$

Then the method is defined by the familiar Strang's splitting recipe:

[21] More precisely it is customary to insist in the integrator being *symplectic* [35] Chapter 6; symplecticness implies conservation of volume and satisfactory —but not exact— conservation of energy, [35], Section 10.3.3. The Verlet scheme is symplectic.

$$\psi_{\Delta t} := \Phi^{(1)}_{\Delta t/2} \circ \Phi^{(2)}_{\Delta t} \circ \Phi^{(1)}_{\Delta t/2}$$

($\Phi^{(i)}$ is the flow of $H^{(i)}$). In this way, given (p^m, x^m), we compute the approximation $(p^{m+1}, x^{m+1}) = \psi_{\Delta t}(p^m, x^m)$ at the next time level by means of the three fractional steps:

$$\begin{aligned} p^{m+1/2} &= p^m - \frac{\Delta t}{2}\frac{\partial}{\partial x}V(x^m), \\ x^{m+1} &= x^m + \Delta t M^{-1} p^{m+1/2}, \\ p^{m+1} &= p^{m+1/2} - \frac{\Delta t}{2}\frac{\partial}{\partial x}V(x^{m+1}). \end{aligned}$$

Since the individual transformations

$$\begin{gathered} (p^m, x^m) \mapsto (p^{m+1/2}, x^m), \\ (p^{m+1/2}, x^m) \mapsto (p^{m+1/2}, x^{m+1}), \\ (p^{m+1/2}, x^{m+1}) \mapsto (p^{m+1/2}, x^{m+1}) \end{gathered}$$

are flows of canonical systems they preserve volume. As a result $\psi_{\Delta t}$ (which is the composition of the three) and $\Psi_T = \psi_{\Delta t}^{T/\Delta t}$ preserve volume.

The reversibility of $\psi_{\Delta t}$ (and hence that of Ψ_T i.e. $S \circ \Psi_T = \Psi_T^{-1} \circ S$) is easily checked and is a consequence of the symmetric pattern of the Strang splitting.

More sophisticated reversible, volume preserving splitting algorithms exist, but the Verlet method is commonly used in molecular simulations and other application areas.

9 The Hybrid Monte Carlo Method

The Hybrid Monte Carlo (HMC) algorithm originated in the physics literature [8] and, while it may be used in other application fields such as Bayesian statistics (see e.g. [14]), its description requires to think of the given problem in physical terms. Let us first present the idea that underlies the method.

9.1 The Idea

Without loss of generality, we write the target density $\pi(x)$ in the state space $\mathbb{R}^d$ as $\exp(-V(x))$ and, regardless of the application in mind, think of $x \in \mathbb{R}^d$ as specifying the configuration of a mechanical system and of $V(x)$ as the corresponding potential energy. We choose arbitrarily $T > 0$ and a positive definite symmetric matrix M (M is often diagonal). Next we consider the Hamiltonian function

$$H = \frac{1}{2} p^T M^{-1} p + V(x)$$

and think of p as momenta and of M as a mass matrix. For the canonical probability distribution in the phase space $\mathbb{R}^D$, $D = 2d$, with density (we set $\beta = 1$ for simplicity)

$$\propto \exp(-H) = \exp\Big(-\frac{1}{2} p^T M^{-1} p\Big) \times \exp\big(-V(x)\big)$$

the random vectors $p \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$ are stochastically independent (we found a similar independence in Section 8.3.1). The (marginal) distribution of x is our target $\pi(x) = \exp(-V(x))$ and p has a Gaussian density $\propto \exp\big(-(1/2) p^T M^{-1} p\big)$ so that samples from p are easily available. In this set-up, Theorem 7 suggests a means to construct a Markov chain in $\mathbb{R}^d$ reversible with respect to the target $\pi(x)$:

Theorem 8. *Define the transitions $x_n \mapsto x_{n+1}$ in the state space $\mathbb{R}^d$ by the following procedure:*

- *Draw p_n from the Gaussian density $\propto \exp\big(-(1/2) p^T M^{-1} p\big)$.*
- *Find $(p^*_{n+1}, x_{n+1}) = \Phi_T(p_n, x_n)$, where Φ_T is the T-flow of the canonical system (26) with Hamiltonian function H.*

Then $x_n \mapsto x_{n+1}$ defines a Markov chain in $\mathbb{R}^d$ that has the target $\pi(x) \propto \exp(-V(x))$ as an invariant probability distribution. Furthermore this Markov chain is reversible with respect to $\pi(x)$.

Proof. The Markov property is obvious: the past enters the computation of x_{n+1} only through the knowledge of x_n. If X_n is distributed $\sim \pi(x)$, then, by the choice of p_n, the random vector (P_n, X_n) has the canonical density $\propto \exp(-H)$. By Theorem 7, (P^*_{n+1}, X_{n+1}) also has density $\propto \exp(-H)$; accordingly, the density of X_{n+1} will be the marginal $\pi(x)$. The reversibility of the chain is a simple consequence of the reversibility of the flow Φ_T. □

The main appeal of this procedure is that, unlike the situation in RW or MALA, the transitions $x_n \mapsto x_{n+1}$ are non-local in the state space $\mathbb{R}^d$, in the sense that x_{n+1} may be far away from the previous state x_n. Fig. 13, as we know, corresponds to the double well potential V in (27) with minima at $x = \pm 1$, so that the target $\exp(-V(x))$ has modes (locations of maximum probability density) at $x = \pm 1$. If the current location x_n is the abscissa of the circles in that figure and the drawing of p_n leads to the point (p_n, x_n) depicted by the upper circle, then the T-flow of the Hamiltonian system yields the upper diamond and x_{n+1} will be the corresponding abscissa. In this way the procedure has carried out, in a single step of the Markov chain, a transition from the neighborhood of the mode at $x = 1$ to the neighborhood of the mode at $x = -1$.

Note that, once x_{n+1} has been determined, the momentum vector p^*_{n+1} is discarded and a fresh p_{n+1} is drawn. Therefore the next starting location (p_{n+1}, x_{n+1}) will have $H(p_{n+1}, x_{n+1}) \neq H(p^*_{n+1}, x_{n+1}) = H(p_n, x_n)$. This makes it possible to explore the whole phase space in spite of the fact that each point only flows within the corresponding level set of the energy H.

Unfortunately the procedure in Theorem 8 cannot be implemented in practice: the required flow Φ_T is not explicitly known except in simple academic examples!

9.2 The Algorithm

In order to turn the procedure we have studied into a practical algorithm, the exact flow Φ_T is replaced by a numerical approximation Ψ_T as in Section 8.4 and an accept/reject mechanism is introduced to assure that the resulting chain still has the target as an invariant distribution. The accep/reject recipe is greatly simplified if integrator Ψ_T is *volume preserving and reversible*, something we assume hereafter (Verlet is integrator of choice).

The transition $x_n \mapsto x_{n+1}$ in HMC is as follows:

- Draw a value p_n from the density $\exp\big(-(1/2)p^T M^{-1} p\big)$.
- Find $(p^*_{n+1}, x^*_{n+1}) = \Psi_T(p_n, x_n)$ (i.e. perform $T/\Delta t$ time-steps of the chosen numerical integrator with step-length Δt). Discard p^*_{n+1} and take x^*_{n+1} as proposal.
- Set $x_{n+1} = x^*_{n+1}$ with probability

$$1 \wedge \exp\Big(-\big(H(p^*_{n+1}, x^*_{n+1}) - H(p_n, x_n)\big)\Big)$$

(acceptance). If the proposal is rejected set $x_{n+1} = x_n$.

Analyses of HMC are given in [7] and [38]. The following result, whose proof is postponed to Section 9.3, holds:

Theorem 9. *In the situation just described, the transitions $x_n \mapsto x_{n+1}$ define a Markov chain reversible with respect to the target $\pi(x) \propto \exp(-V(x))$.*

Some comments are in order. If the exact flow Φ_T were known and we used it as 'numerical integrator', i.e. $\Psi_T = \Phi_T$, then, by conservation of energy,

$$\exp\Big(-\big(H(p^*_{n+1}, x^*_{n+1}) - H(p_n, x_n)\big)\Big) = 1$$

and every proposal would be accepted: one is then back in the procedure covered by Theorem 8. In a similar vein, the better the numerical scheme Ψ_T preserves H the higher the probability of acceptance.[22]

With HMC, a *Markov chain step* $x_n \mapsto x_{n+1}$ requires $T/\Delta t$ *time-steps* of the numerical integrator. In the particular case where the integrator is the Verlet scheme and $\Delta t = T$, so that there is a single time-step per step of the chain, it is easy to check that HMC is identical to MALA with $h = \Delta t$ (more precisely, given x_n, the proposal x^*_{n+1} and the accept/reject mechanism are the same in both algorithms).

[22] In this connection it may be worth noting that the proof of Theorem 9 demands that the mapping Ψ_T is time reversible and volume preserving, but would work even if Ψ_T were not an approximation to the true Φ_T. However if Ψ_T is not close to Φ_T, the acceptance probability will be low.

This equivalence is somewhat surprising as MALA proposals are motivated by an SDE, whereas HMC proposals are based on deterministic Hamiltonian dynamics.[23] After this equivalence MALA/HMC one may think of HMC as a non-local version of MALA.

The paper [3] shows that if the target consists of d independent copies of the same distribution, then the Verlet time-step Δt should be chosen $\propto (1/d)^{1/4}$ to have $\mathcal{O}(1)$ acceptance probabilities as $d \to \infty$. For reversible, volume preserving integrators of (necessarily even) order 2ν, $\Delta t \propto (1/d)^{1/(4\nu)}$. This compares favorably with the corresponding relations for RW and MALA reported in Section 7.

In our description in Section 8.4 we observed that the Verlet algorithm is based on splitting H into its kinetic and potential part. This is not the only possibility of splitting, see [37], [4]. Modifications of HMC may be seen in [1], [20] and [19], among others.

We now turn our attention to an example. As in Section 7, consider the target $\propto \exp(-(1/2)k(r-1)^2)$, $k = 100$ but now $d = 3$. We show the draws in the two-dimensional plane of the random vector (x_1, x_2) (the corresponding marginal is approximately uniform on the unit disk) and the correlation in the variable x_1. Figs. 14, 15 and 16 show results for RW, MALA and HMC (with Verlet integration) respectively. The superiority of HMC in this example is manifest.

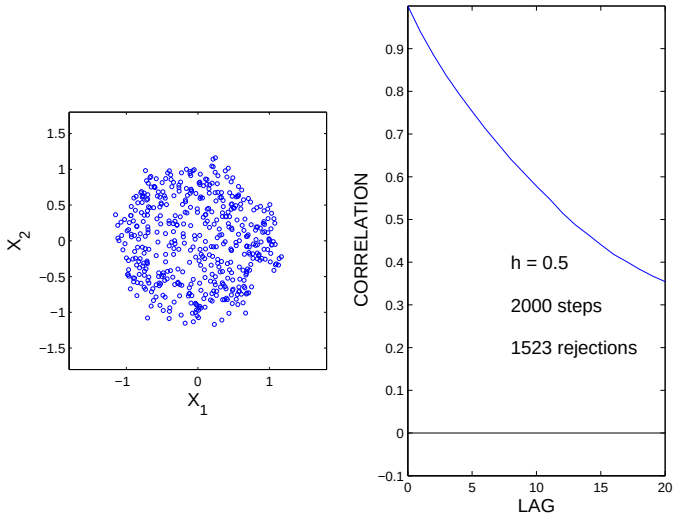

Fig. 14 RW results for a stiff spring in $\mathbb{R}^3$. Samples of the coordinates x_1, x_2 and autocorrelation in x_1

[23] Recall that when the MALA proposal is seen as an Euler-Maruyama step for an SDE, the MALA parameter h coincides with the *square* of the time-step Δt. However in the relation of MALA with HMC studied in this section, $h = \Delta t$ as we have just pointed out.

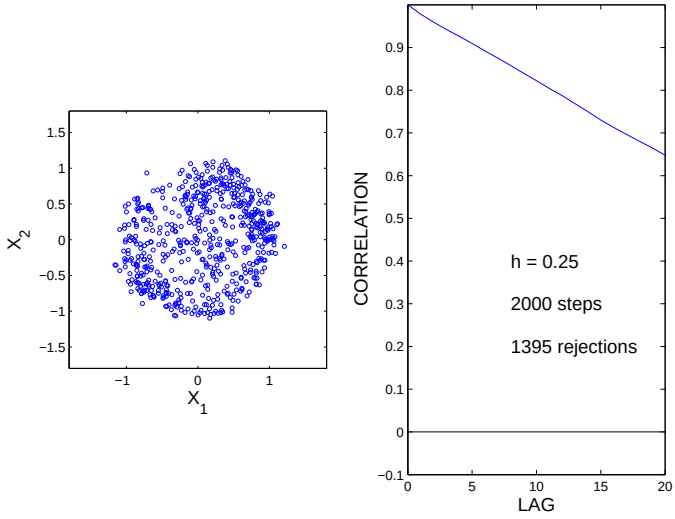

Fig. 15 MALA results for a stiff spring in $\mathbb{R}^3$. Samples of the coordinates x_1, x_2 and autocorrelation in x_1

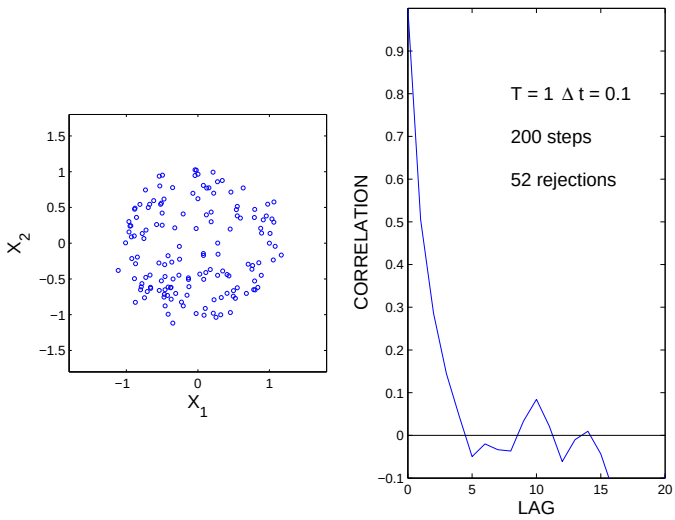

Fig. 16 HMC results for a stiff spring in $\mathbb{R}^3$. Samples of the coordinates x_1, x_2 and autocorrelation in x_1

9.3 Proofs

The proof of Theorem 9, borrowed from [23], is based on some lemmas. We shall use repeatedly the fact that the momentum-flip symmetry S, $S(p,x) = (-p,x)$, preserves the canonical probability measure μ: $\mu(S(A)) = \mu(A)$ for each Borel subset A of $\mathbb{R}^D$.

Lemma 1. *Consider a Borel probability measure μ in $\mathbb{R}^D$ that is preserved by the momentum-flip symmetry S and a transition kernel K in the phase space $\mathbb{R}^D$ that satisfies the following analogue of the detailed balance condition (14):*

$$\int_A \mu(d\xi)\, K(\xi, B) = \int_B \mu(d\eta)\, K(S(\eta), S(A)) \tag{29}$$

for each (Borel measurable) A, B. Then (cf. Theorem 3):

- *The measure μ is invariant with respect to K.*
- *At stationarity, the chain $\Xi_0, \ldots, \Xi_N$ generated by K is statistically the same as the chain $S(\Xi_N), \ldots, S(\Xi_0)$.*

Proof. With $A = \mathbb{R}^D$, the hypothesis (29) implies,

$$\int_{\mathbb{R}^D} \mu(d\xi) K(\xi, B) = \int_B \mu(d\eta) K(S(\eta), \mathbb{R}^D) = \int_B \mu(d\eta) = \mu(B);$$

this proves stationarity (see (13)).

By definition of conditional probability,

$$\begin{aligned} &\mathbb{P}\big(S(\Xi_n) \in S(A) \mid S(\Xi_{n+1}) \in S(B)\big) \\ &\qquad = \mathbb{P}\big(\Xi_n \in A \mid \Xi_{n+1} \in B\big) = \frac{\mathbb{P}\big(\Xi_n \in A \wedge \Xi_{n+1} \in B\big)}{\mathbb{P}\big(\Xi_{n+1} \in B\big)}. \end{aligned} \tag{30}$$

Let us rewrite the last fraction. At stationarity $\mathbb{P}\big(\Xi_{n+1} \in B\big) = \mathbb{P}\big(\Xi_n \in B\big)$; furthermore, after the change of variables $\eta = S(\xi)$, (29) becomes (recall that $\mu\big(S(d\xi)\big) = \mu(d\xi)$)

$$\int_A \mu(d\xi)\, K(\xi, B) = \int_{S(B)} \mu(d\xi)\, K(\xi, S(A)),$$

which means

$$\mathbb{P}\big(\Xi_n \in A \wedge \Xi_{n+1} \in B\big) = \mathbb{P}\big(\Xi_n \in S(B) \wedge \Xi_{n+1} \in S(A)\big)$$

Taking these results back to (30)

$$\begin{aligned} &\mathbb{P}\big(S(\Xi_n) \in S(A) \mid S(\Xi_{n+1}) \in S(B)\big) \\ &\qquad = \frac{\mathbb{P}\big(\Xi_n \in S(B) \wedge \Xi_{n+1} \in S(A)\big)}{\mathbb{P}\big(\Xi_n \in S(B)\big)} = \mathbb{P}\big(\Xi_{n+1} \in S(A) \mid \Xi_n \in S(B)\big). \end{aligned}$$

□

Lemma 2. *As above, let μ be a measure preserved by the momentum-flip map S. Assume that K^* is a (proposal) Markov kernel in $\mathbb{R}^D$, such that the measures*

$$K^*\big(S(\eta), S(d\xi)\big)\,\mu(d\eta), \qquad K^*(\xi, d\eta)\,\mu(d\xi)$$

in $\mathbb{R}^D \times \mathbb{R}^D$ are equivalent (i.e. each has a density with respect to the other), so that there is a function r such that

$$r(\xi,\eta) = \frac{K^*\big(S(\eta), S(d\xi)\big)\,\mu(d\eta)}{K^*(\xi, d\eta)\,\mu(d\xi)}. \tag{31}$$

Define a Markov transition $\xi_n \mapsto \xi_{n+1}$ in $\mathbb{R}^D$ by:

- *Propose ξ^*_{n+1} according to $K^*(\xi_n, \cdot)$.*
- *Accept ($\xi_{n+1} = \xi^*_{n+1}$) with probability $1 \wedge r(\xi_n, \xi^*_{n+1})$. If rejection occurs set $\xi_{n+1} = S(\xi_n)$.*

The chain defined in this way satisfies the generalized detailed balance condition (29) and, in particular, μ is an invariant measure.

Proof. The kernel of the chain is:

$$K(\xi, d\eta) = \big(1 \wedge r(\xi,\eta)\big)\, K^*(\xi, d\eta) + \big(1 - \alpha(\xi)\big)\, \delta_{S(\xi)}(d\eta),$$

where

$$\alpha(\xi) = \int_{\mathbb{R}^D} \big(1 \wedge r(\xi,\eta)\big)\, K^*(\xi, d\eta)$$

is the probability of acceptance conditioned to $\Xi_n = \xi$ and δ denotes a point unit mass (Dirac's delta).

We have then to show that

$$\begin{aligned} &\big(1 \wedge r(\xi,\eta)\big)\, K^*(\xi, d\eta)\,\mu(d\xi) + \big(1 - \alpha(\xi)\big)\, \delta_{S(\xi)}(d\eta)\,\mu(d\xi) = \\ &\big(1 \wedge r(S(\eta), S(\xi))\big)\, K^*\big(S(\eta), S(d\xi)\big)\,\mu(d\eta) + \big(1 - \alpha(S(\eta))\big)\, \delta_\eta(S(d\xi))\,\mu(d\eta), \end{aligned} \tag{32}$$

a task that we carry out by proving that the first and second term in the left-hand side coincide with the first and second term in the right-hand side respectively. For the second terms, if ϕ is a test function, the change of variables $\xi = S(\xi')$ enables us to write

$$\begin{aligned} &\int_{\mathbb{R}^D \times \mathbb{R}^D} \phi(\xi,\eta)\,\big(1 - \alpha(S(\eta))\big)\,\delta_\eta(S(d\xi))\,\mu(d\eta) = \\ &\qquad \int_{\mathbb{R}^D \times \mathbb{R}^D} \phi(S(\xi'),\eta)\big(1 - \alpha(S(\eta))\big)\,\delta_\eta(d\xi')\,\mu(d\eta) \end{aligned}$$

and, by definition of δ, the last integral has the value

$$\int_{\mathbb{R}^D} \phi(S(\eta),\eta)\big(1 - \alpha(S(\eta))\big)\,\mu(d\eta).$$

Now the change of variables $\eta = S(\xi)$ and the definition of δ allow us to continue

$$\int_{\mathbb{R}^D} \phi(S(\eta),\eta)\,\big(1-\alpha(S(\eta))\big)\,\mu(d\eta) = \int_{\mathbb{R}^D} \phi(\xi,S(\xi))\,\big(1-\alpha(\xi)\big)\,\mu(d\xi)$$
$$= \int_{\mathbb{R}^D\times\mathbb{R}^D} \phi(\xi,\eta)\,\big(1-\alpha(\xi)\big)\,\delta_{S(\xi)}(d\eta)\,\mu(d\xi).$$

This proves that the second terms in (32) are equal. For the first terms note that $r(\xi,\eta) = 1/r\big(S(\eta),S(\xi)\big)$. Thus:

$$\big(1\wedge r(\xi,\eta)\big)\,K^*(\xi,d\eta)\,\mu(d\xi) = \big(r(S(\eta),S(\xi))\wedge 1\big)\,r(\xi,\eta)\,K^*(\xi,d\eta)\mu(d\xi)$$

and by definition of r this has the value:

$$\big(1\wedge r(S(\eta),S(\xi)\big)\,K^*\big(S(\eta),S(d\xi)\big)\,\mu(d\eta).$$

□

Lemma 3. *Let μ be the measure in $\mathbb{R}^D$ with density* $\exp(-H)$*, where $H\circ S = H$ and assume that Ψ_T is a reversible and volume preserving transformation in phase space (in particular the numerical solution operator associated with a reversible, volume preserving integrator for the Hamiltonian system associated with H). Define a transition kernel by*

$$K^*(\xi,d\eta) = \delta_{\Psi_T(\xi)}(d\eta).$$

Then μ and K^* satisfy the requirements in Lemma 2 and the Metropolis-Hastings ratio r in (31) has the value $\exp\big(-(H(\Psi_T(\xi))-H(\xi))\big)$.

Proof. For the measure in the numerator of (31), the integral of a test function ϕ is

$$I_N = \int_{\mathbb{R}^D\times\mathbb{R}^D} \phi(\xi,\eta)\,\delta_{\Psi_T(S(\eta))}(S(d\xi))\,\mu(d\eta).$$

Changing $\xi = S(\xi')$ leads to

$$I_N = \int_{\mathbb{R}^D\times\mathbb{R}^D} \phi(S(\xi'),\eta)\,\delta_{\Psi_T(S(\eta))}(d\xi')\,\mu(d\eta)$$
$$= \int_{\mathbb{R}^D} \phi\big(S(\Psi_T(S(\eta))),\eta\big)\,\mu(d\eta).$$

Now use the reversibility of Ψ_T and the definition of μ to write

$$I_N = \int_{\mathbb{R}^D} \phi(\Psi_T^{-1}(\eta),\eta)\,\exp(-H(\eta))\,d\eta$$

and then change $\eta = \Psi_T(\xi)$

$$I_N = \int_{\mathbb{R}^D} \phi(\xi,\Psi_T(\xi))\,\exp(-H(\Psi_T(\xi)))\,d\xi.$$

(Note we have used here conservation of volume.)

The integral with respect to measure in denominator of (31) is:

$$I_D = \int_{\mathbb{R}^D \times \mathbb{R}^D} \phi(\xi,\eta)\, \delta_{\Psi_T(\xi))}(d\eta)\, \mu(d\xi) = \int_{\mathbb{R}^D} \phi(\xi, \Psi_T(\xi))\, \exp(-H(\xi))\, d\xi$$

and a comparison with I_N leads to the sought conclusion. □

After these preparations we may present the proof of Theorem 9.

Proof. Consider the chain $\mathscr{C}$ in the phase space $\mathbb{R}^D$ of the variable (p,x) such that one step $(p_n,x_n) \mapsto (p_{n+1},x_{n+1})$ of $\mathscr{C}$ is the concatenation of two sub-steps:

1. Discard the value of the momentum p_n and replace it by a fresh sample from the Maxwell distribution for the momentum.
2. Take a step of the chain defined in Lemmas 2 and 3 .

Both sub-steps preserve μ (for the second use Lemma 2) and as a consequence so does the chain $\mathscr{C}$. The x-marginal of $\mathscr{C}$ is the chain in the HMC algorithm and will preserve the marginal density $\exp(-V(x))$. □

Acknowledgements This work has been supported by Project MTM2010-18246-C03-01, Ministerio de Ciencia e Innovación, Spain.

References

1. Akhmatskaya, E., Reich, S.: GSHMC: An efficient method for molecular simulations, J. Comput. Phys. **227**, 4934–4954 (2008)
2. Arnold, V.I.: Mathematical Methods of Classical Mechanics (2nd edn.). Springer, New York (1989)
3. Beskos, A., Pillai, N., Roberts, G.O., Sanz-Serna, J.M., Stuart, A.M.: Optimal tuning of the Hybrid Monte-Carlo Algorithm, Bernoulli (to appear).
4. Beskos, A., Pinski, F.J., Sanz-Serna, J.M., Stuart, A.M.: Hybrid Monte-Carlo on Hilbert spaces, Stoch. Proc. Appl. **121**, 2201-2230 (2011)
5. Billingsley, P.: Probability and Measure (3rd edn). Wiley, New York (1995)
6. Brémaud, P.: Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues. Springer, Berlin (1999)
7. Cancès, E., Legoll, F., Stoltz, G.: Theoretical and numerical comparison of some sampling methods for molecular dynamics, ESAIM: M2AN **41**, 351–389 (2007)
8. Duane, S., Kennedy, A.D., Pendleton, B., Roweth, R.: Hybrid Monte Carlo. Phys. Lett. B **195**, 216-222 (1987)
9. Feller, W.: An Introduction to Probability Theory and Its Applications. Vol 1 (3rd edn). Wiley, New York (1968)
10. Feller, W.: An Introduction to Probability Theory and Its Applications. Vol 1 (2nd edn). Wiley, New York (1971)
11. Feynman, R.P., Leighton, R.B., Sands, M.: The Feynman Lectures on Physics, Vol. 1. Addison-Wesley, Reading (1963)
12. Friedman, A.: Stochastic Differential Equations and Applications. Dover, Mineola N.Y. (2006)
13. Geyer, C.J.: Practical Markov Chain Monte Carlo. Statist. Sci. **7**, 473–483 (1992)
14. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Statist. Soc. B **73**, 123–214 (2011)

15. Grimmett, G., Stirzaker, D.: Probability and Random Processes (3rd edn). Oxford University Press, Oxford (2001)
16. Hairer, E., Lubich, Ch., Wanner, G.: Geometric Numerical Integration (2nd edn). Springer, Berlin (2006)
17. Hastings, W.: Monte Carlo sampling methods using Markov chains and their application. Biometrika **57**, 97–109 (1970)
18. Higham, D.: An algorithmic introduction to numerical simulation of stochastic differential equations, SIAM Rev. **4**3, 525–546 (2001)
19. Hoffman, M.D., Gelman, A.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, preprint.
20. Izaguirre, J.A., Hampton, S.S.: Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules, J. Comput. Phys. **200**, 581–604 (2004)
21. Lawrie, I.D.: A Unified Grand Tour of Theoretical Physics. Institute of Physics Publishing, Bristol (1990)
22. Leimkuhler, B., Reich, S.: Simulating Hamiltonian dynamics. Cambridge University Press, Cambridge (2004)
23. Lelievre, T., Rousset, M., Stoltz, G.: Free Energy Computations: A Mathematical Perspective. Imperial College Press, London (2010)
24. Mao, X.: Stochastic Differential Equations and Applications (2nd edn). Horwood, Chichester (2008)
25. Mattingly, J.C., Pillai, N.S., Stuart, A.M.: Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions Authors. Ann. Appl. Prob. (to appear).
26. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
27. Meyn, S., Tweedie, R.: Markov Chains and Stochastic Stability. Springer, New York (1993)
28. Neal, R.: Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto (1993)
29. Pillai, N.S., Stuart, A.M., Thiery, A.H.: Optimal Scaling and Diffusion Limits for the Langevin Algorithm in High Dimensions. Ann. Appl. Prob. (to appear).
30. Robert, Ch. P., Casella, G.: Monte Carlo Statistical Methods (2nd ed). Springer, Berlin (2004)
31. Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin diffusions and their discrete approximations. Bernoulli **2**, 341–263 (1996)
32. Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**, 110–120 (1997)
33. Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. J. R. Statist. Soc. B **60**, 255–268 (1998)
34. Sanz-Serna, J.M.: Geometric integration. In: Duff, I.S., Watson, A.G. (eds.) The State of the Art in Numerical Analysis, pp. 121–143. Clarendon, Oxford (1997)
35. Sanz-Serna, J.M., Calvo, M.P.: Numerical Hamiltonian Problems. Chapman & Hall, London (1994)
36. Schilick, T.: Molecular Modeling and Simulation: An Interdisciplinary Guide (2nd ed). Springer, New York (2010)
37. Shahbaba, B., Lan, S., Johnson, W.O., Neal, R.M.: Split Hamiltonian Monte Carlo, preprint.
38. Schütte, C.: Conformational Dynamics: Modelling, Theory, Algorithmm and Application to Biomolecules. Habilitation Thesis, Free University Berlin (1999)