Monte Carlo Methods for Sampling the Potential Energy Landscape Ensemble

Vale Cofer-Shabica

Modified: October 23, 2014 at 15:10

Abstract

I discuss the general scheme and background for Monte Carlo random walks. This is then applied this to the specific case of sampling the "roaming region" of the formaldehyde (H_2CO) potential energy surface.

1 Introduction

*The first part of this section is taken **directly** from [1], whose methods I have poached.*

Metropolis sampling is a Monte Carlo procedure for evaluating ensemble averages of the form:

$$\langle f \rangle = \frac{\int dx^N P(\vec{x}) f(\vec{x})}{\int dx^N P(\vec{x})} \tag{1}$$

where f is a property of the N-dimensional space determined by $\vec{x}$ and $P(\vec{x})$ is the probability of a given state, $\vec{x}$. The Metropolis algorithm is of particular use when $P(\vec{x})$ is difficult to sample directly or to normalize. Its output is a set of states, $\{\vec{x}_i\}$, such that the average in eq. 1 can be replaced by:

$$\langle f \rangle = \frac{1}{N} \sum_i f(\vec{x}_i) \tag{2}$$

The set, $\{\vec{x}_i\}$, is generated by a biased random walk in the space which compares the relative probabilities, P, for the current and succeeding states. In this way the probability of being in a particular state at a particular step only depends explicitly on the prior step; such a sequence is called a "Markov chain".

I seek to generate such a set which samples the "roaming region" of the formaldehyde potential energy surface. The set will then be used to compute geodesics through the region.

1.1 Details

Given an initial state, $\vec{x}_0$, a (potentially un-normalized) probability distribution, $P(\vec{x})$, and a propagator for generating *trial* moves, $F : \vec{x}_n \to \vec{x}'_{n+1}$, then the sequence of the random walk

is given recursively as follows:

$$\vec{x}_0 = \vec{x}_0$$
$$\vec{x}_{n+1} = \begin{cases} F(\vec{x}_n) & \text{, with probability: } \min\left[1, \frac{P(\vec{x}'_{n+1})}{P(\vec{x}_n)}\right] \\ \vec{x}_n & \text{, otherwise} \end{cases} \tag{3}$$

The probabilistic component of eq. 3 can be handled by comparing a uniform, random variable[1] on $[0, 1)$, ξ, to the target probability, p, and accepting if $\xi < p$. If this is confusing, think of the limiting cases: $p = 0$ and $p = 1$.

F, the propagator can take many forms, but the one used in Brady *et al.* [1] and here is the following:[2]

$$F(\vec{x}) = \vec{x} + \hat{x}_i(\xi - 0.5)\delta_x \tag{4}$$

where $\hat{x}_i$ is a unit vector along a randomly chosen coordinate, ξ is defined as above, and δ_x is a scaling factor. In the case of sampling configuration space, this amounts to is moving a single random atom, along a random coordinate, by a random amount.

There is a bit of lore surrounding Markov chains, which indicates that an acceptance ratio (*i.e.*, the fraction of accepted trial steps) of 50% is "best". Conversations with Jimmie Doll indicated that this is very rough and that a better guide is perhaps "bigger than 10% and less than 90%". One can adjust the scaling factor, δ_x, to achieve the desired acceptance ratio during a trial phase. However, once data-collection has begun (once "real" data is being collected), δ_x must remain fixed or the chain will fail to satisfy detailed balance and the intended distribution will not be sampled. I describe a naïve implementation of a scheme to pick δ_x in section 2.2.

2 Methods

I now describe the specifics of my implementation for sampling points from the "roaming region" of the formaldehyde potential energy surface.

2.1 A Probability-like Function

In the first section I stated that we could use un-normalized probability distributions. This is because the distribution, P only appears in eq. 3 in ratio with itself. This is quite convenient because it obviates the need to compute what would be—in this case and many others—a complicated, many-dimensional configurational integral.

My goal is to sample the "roaming region" of the formaldehyde PES. This is a subset of the total configuration space of formaldehyde, which I specify as the 12 Cartesian coordinates of its 4 atomic centers. Per my own definition, a point in this space is in the roaming region if the following criteria are satisfied:

[1] All random numbers were generated using the Mersenne Twister as implemented in the GNU Scientific Library [2].

[2] In the course of the discussion that ensued in group meeting, issues with center-of-mass creep were rightly raised. The given propagator does *not* preserve the center of mass. This is inappropriate as the boundary conditions for the geodesic algorithm should have the same center of mass (This is because we have field-free, translationally-invariant potential.). To correct this, an additional step is taken after the generation of $F(\vec{x}_n)$. The center of mass of the system is reset by subtracting $(\vec{x}_{cm} - \vec{x}_{target})$ from the coordinates of each center, where $\vec{x}_{cm}$ is the center of mass of $F(\vec{x}_n)$ and $\vec{x}_{target}$ is the desired center of mass.

1. The restoring force on the roaming hydrogen is minimally attractive: $F_{H^{(0)}} > F_{min}$
2. The hydrogens are separated by more than a minimum distance: $\|\vec{r}_{HH^{(0)}}\| > d_{min}$
3. The hydrogens are separated by less than a maximum distance: $\|\vec{r}_{HH^{(0)}}\| < d_{max}$

The roaming hydrogen, designated $H^{(0)}$, is uniquely identified in configuration space as the hydrogen with the greatest Euclidean separation from total center of mass. I define the restoring force on the roaming hydrogen as:

$$F_{H^{(0)}} = -\hat{r}_{H^{(0)}-CM} \cdot \vec{\nabla}_{H^{(0)}} V(\vec{R}) \tag{5}$$

where

$$\vec{\nabla}_{H^{(0)}} = \left(\frac{\partial}{\partial x_{H^{(0)}}}, \frac{\partial}{\partial y_{H^{(0)}}}, \frac{\partial}{\partial z_{H^{(0)}}} \right) \tag{6}$$

$$\vec{r}_{H^{(0)}-CM} = \vec{r}_{H^{(0)}} - \vec{r}_{CM} \tag{7}$$

and $V(\vec{R})$ is the formaldehyde potential energy function; $\hat{r}_{H^{(0)}-CM}$ is a unit vector defined in the usual way.

The constants in the criteria are given below and will be justified in a forthcoming appendix.

Parameter	Value
F_{min}	$-2.5 \times 10^{-4}\ E_H/a_0$
d_{min}	5.86728 a_0
d_{max}	9.00000 a_0

Briefly, however:

1. $F_{H^{(0)}}$ is the force on the roaming hydrogen towards the center of mass. As usual, negative values are attractive. I place a minimum on the force because in roaming trajectories, we observe large, persistent H - HCO separations, which require low attractive forces on the wayfaring hydrogen.
2. A minimum separation between the hydrogens in required because otherwise equilibrium and transition state configurations, clearly not representative of roaming, would be mis-classified as roaming. I use the hydrogen separation as opposed to H - HCO separation because large hydrogen - formyl separation can also correspond to dissociation to molecular products.
3. A maximum hydrogen separation is also imposed to exclude radical dissociation, which otherwise comprises a majority of the space.

These criteria can be formally encoded in a function as follows:

$$P_{roaming}(\vec{R}; F_{min}, d_{min}, d_{max}) \propto \Theta\left(F_{H^{(0)}} - F_{min}\right) \tag{8}$$
$$\cdot\Theta\left(\|\vec{r}_{HH^{(0)}}\| - d_{min}\right) \tag{9}$$
$$\cdot\Theta\left(d_{max} - \|\vec{r}_{HH^{(0)}}\|\right) \tag{10}$$

where Θ is the step function:

$$\Theta(x) = \begin{cases} 1 & , x > 0 \\ 0 & , \text{otherwise} \end{cases} \tag{11}$$

To sample the potential energy landscape ensemble [3], we can use a similar scheme:

$$P_{pele}(\vec{R}; E_L) \propto \Theta\left(E_L - V(\vec{R})\right) \tag{12}$$

where E_L is the landscape energy.

Combining equation 12 with 8 allows us to sample the portion of the roaming region that is also in a given potential energy landscape ensemble. This is desirable because it allows us to set the landscape energy for sampled configurations in advance. Perhaps this is an element of the technique that would be useful to other members of the group.

It should be noted that since our combined function has binary outputs, the probabilities in eq. 3 will always be 0 or 1. This is a special case of the more general class of problems than the Metropolis algorithm is capable of handling—within the allowed region, we seek uniform sampling rather than a variable density.

Put another way, because we have no notion of "sort of roaming" or "sort of in the potential energy landscape ensemble", the random walk will never do something "sort of bad" *i.e.*, will never take "uphill steps". When Brady and co-workers[1] implemented this scheme, it was to sample the energy shell of the microcanonical ensemble. While there is no notion of "sort of the right energy" for the microcanonical ensemble, they used a pre-limit form of the delta function for their probability, which allowed their walker to meander back and forth across the energy shell.

2.2 Picking a Scaling Factor

For the linear propagator described in section 1.1, we would like to optimize the scaling factor, δ_x, such that the acceptance ratio falls within the desired bounds.

The binary probability function could introduce some added complications, but we avoid them under the following assumption:

- The set to be sampled is dense; that is, points in the set are arbitrarily close to others in the set

This condition allows us to assume that for sufficiently small steps, moves will *always* fall within the acceptance region. This is important because it allows us to take the $\delta_x \to 0$ limit as yielding an acceptance ratio of 1. In the case of a binary probability function, the other limit, $\delta \to \infty$, implies an acceptance ratio equal to the value $\langle P(\vec{x})\rangle$. In the case of the roaming region of formaldehyde, this average is 0.

It would also be convenient to assume that the set is connected; that is, it is possible to translate between any two points in the set while remaining in the set. This a statement of the ergodicity of the system and is actually a stronger requirement than the first assumption. It allows us to assume that the entire set is reachable from any x_0. *However*, in the potential energy landscape ensemble, we have no guarantees that this assumption will hold.

The general scheme for selecting δ_x is this:

1. Guess an initial value for δ_x
2. Compute a Markov chain of length k while keeping track of the acceptance ratio
3. If the ratio is larger than desired, scale by $(1+\alpha)$ to increase the length. Likewise, if the ratio is smaller than desired, scale by $(1-\alpha)$ to decrease the step.
4. Goto 2 and repeat M times.

If α is too large, the results will be unstable as δ_x oscillates in and out of the acceptable range. Too small and convergence will take a very long time. Similar observations can be made about k. I found acceptable results with $\alpha = 0.05$ and $k = 1000$. Using this plan for picking δ_x I achieved an acceptance ratio of 82.30 % with a scaling factor of 0.2900 a_0 after 1.1×10^6 total steps. This step size was then use for sampling the roaming region. All of my parameters are given in the table below.

Parameter	Value
δ_x (guess)	0.01 a_0
k	1000
target acceptance ratio	10% to 90%
α	0.05
M	1100
δ_x (final)	0.2900 a_0
acceptance ratio	82.30%

3 Results

The objects in section 1.1 are general and, desiring to exploit this, I implemented them in C in such a way that arbitrary probability functions and propagators could be used[3].

3.1 A simple test

As a test of my implementation, I sampled the set defined by:

$$P(\vec{x}) = \Theta(1 - \|\vec{x}\|) \tag{13}$$

in 2 and 3 dimensions. This is, of course, the unit-ball. I used all the other procedures described in section 2 and obtained an acceptance ratio of 37.18 % in generating a chain of 10^4 steps using a step size of $\delta_x = 4.62884$. A plot of my results in 2 dimensions appears in figure 1

I also computed correct-looking distributions for an annulus in 2 and 3 dimensions:

$$P(\vec{x}) = \Theta\left(1 - \|\vec{x}\|\right) \cdot \Theta\left(\|\vec{x}\| - \frac{1}{2}\right) \tag{14}$$

and a variable-density version of the ball:

$$P(\vec{x}) = \Theta(1 - \|\vec{x}\|) \cdot \|\vec{x}\|^\alpha \tag{15}$$

where $\alpha = 1, 2$ in 2 and 3 dimensions.

[3]This was certainly no more time-consuming than a one-off implementation because it forced me to structure my functions in an organized way.

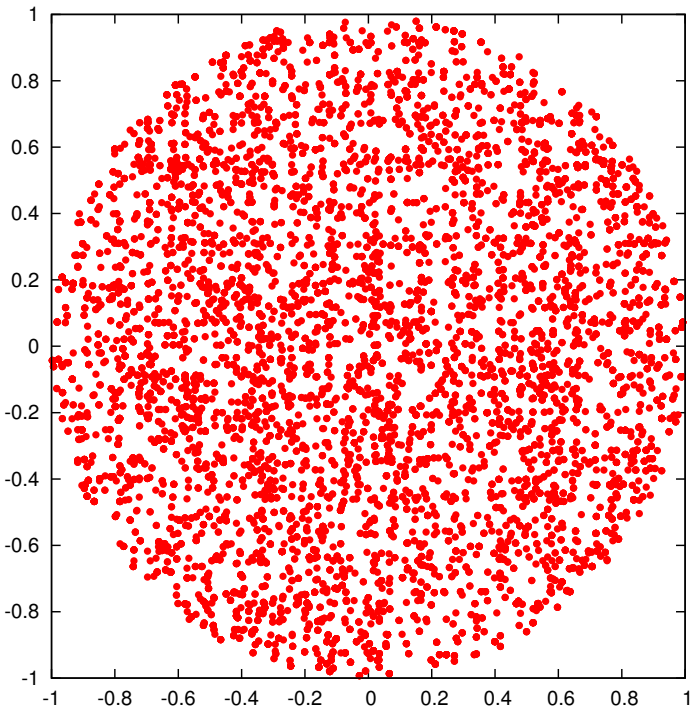

Figure 1: Test of the Markov chain sampling scheme in the unit disk

3.2 The Formaldehyde Potential

Given that there is no reason to expect that the roaming region is self-connected, I wanted to use many seed values for the Markov chain. To this end, I collected all points from all 37600 cm^{-1} MD trajectories subject to the following constraints:

- trajectories terminated within 12.1 ps as molecular products
- the points satisfied $V(\vec{x}) < 0.1591467 E_H$; this was the landscape energy used in my first analysis.
- points were in the "roaming region" as defined in section 2.1

These criteria yielded 1143 points in the roaming region. Using each of the points as a different value for x_0, I generated Markov chains of length 10^5, recording values every 100 moves. This gave a list of 1.143×10^6 points in the roaming region with an average acceptance ratio $83.7742 \pm 0.7272\%$). These points were then shuffled and used as intermediate roaming points in the geodesic path-finding algorithm. A few visualizations of the generated points follow.

The figures show the position of the roaming hydrogen in a reference frame such that the origin is the formyl center of mass, the HCO plane is the xy plane and the CO axis is parallel to the y axis. Red coloration indicates attraction to formyl group while blue, repulsion.

The figures are promising because they seem to indicate that the algorithm is behaving as it should[4]. The case of the single chain depicted in figure 2 shows the random walker beginning to explore a small arc of the region. In figure 3, we see that the region appears to be uniformly sampled.

[4]Yes, this is quite weak, what tests should/can I do to verify their validity?

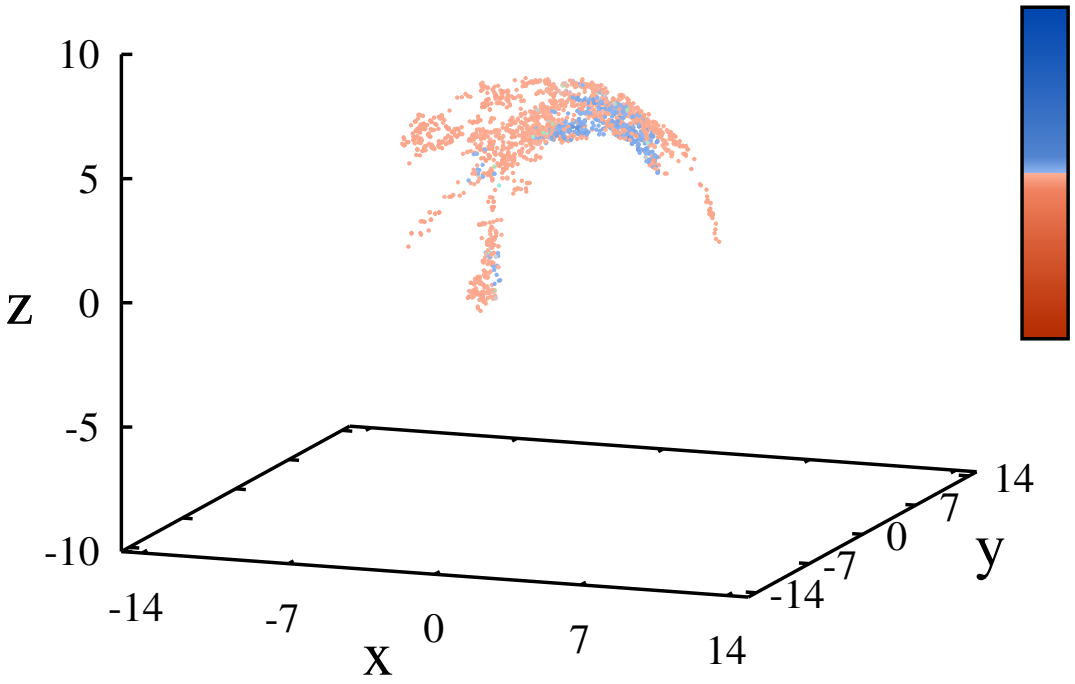

Figure 2: 2.5×10^3 randomly selected points from a single 10^4 state Markov chain on the formaldehyde PES.

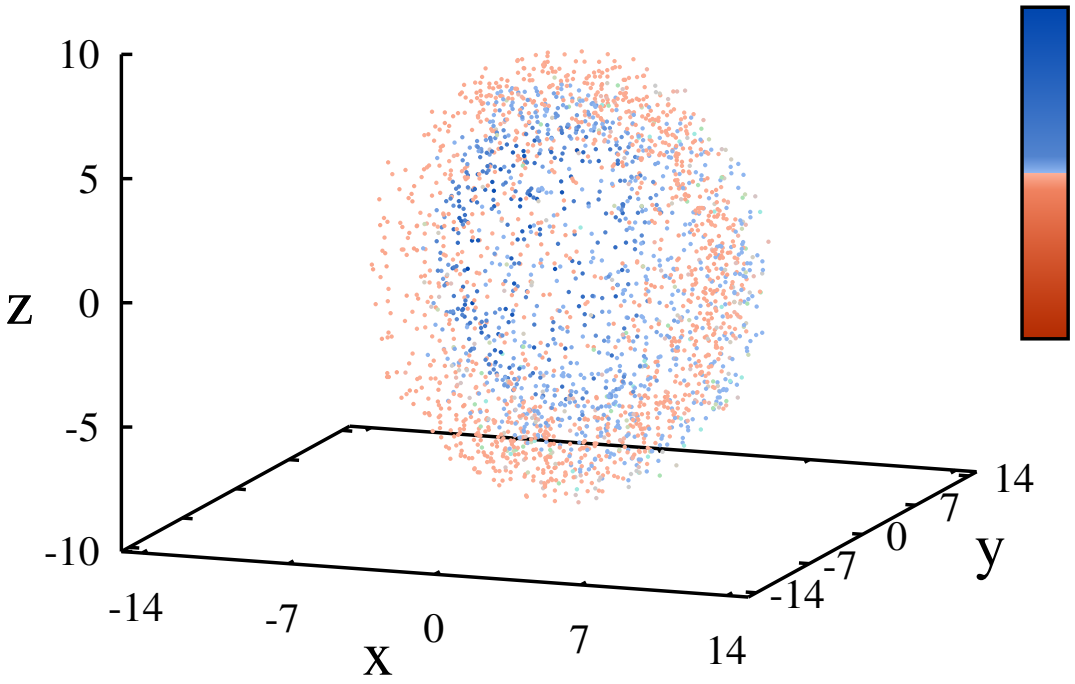

Figure 3: 2.5×10^3 randomly selected points from the 1.14×10^6 roaming points generated on the formaldehyde PES.

References

[1] Brady, J. W.; Doll, J. D.; Thompson, D. L. *The Journal of Chemical Physics* **1981**, *74*, 1026–1028.

[2] Galassi, M.; Davies, J.; Theiler, J.; Gough, B.; Jungman, G.; Alken, P.; Booth, M.; Rossi, F. *GNU Scientific Library Reference Manual*, 3rd ed.; Network Theory, Ltd., 2009; www.gnu.org/software/gsl.

[3] Wang, C.; Stratt, R. M. *The Journal of Chemical Physics* **2007**, *127*, 224503.