Hamiltonian Importance Sampling

Radford M. Neal, University of Toronto

http://www.cs.utoronto.ca/∼radford

BIRS Workshop on Molecular Dynamics, June 2005

Review of Importance Sampling

We want to estimate expectations with respect to the distribution with probability density $\pi(x) = f(x)/Z_f$, where $Z_f = \int f(x)dx$.

Suppose we can't sample from $\pi(x)$. Instead, we sample from the distribution with density proportional to $g(x)$, with normalizing constant $Z_g = \int g(x)dx$.

Given points $x_1, \ldots, x_n$ drawn from g, we can estimate $\langle a \rangle_f$, the expectation of $a(x)$ with respect to π, by

$$\sum_{i=1}^{n} w_i a(x_i) \Big/ \sum_{i=1}^{n} w_i$$

Here, $w_i = f(x_i)/g(x_i)$ is the *importance weight* for point x_i.

We can estimate the ratio Z_f/Z_g by $(1/n) \sum_{i=1}^{n} w_i$.

Difficulties with Importance Sampling

For a complex, high-dimensional distribution $\pi(x)$, it is difficult to choose a distribution $g(x)$ that satisfies all of the following requirements:

1) *It is a good approximation to* π. If not, the importance weights will be highly variable, and the effective sample size when estimating $\langle a \rangle_f$ will be very small.

2) *We can feasibly sample from it (independently).* Easily-sampled distributions like Gaussians aren't good approximations. We need something like the distribution defined by K Metropolis updates starting from a uniformly-distributed start state.

3) *We can compute* $g(x)$*, and hence the importance weights.* Sadly, the density for the distribution defined by K Metropolis updates involves an infeasible integral over all intermediate states.

Jarzynski's Method

Jarzynski's method — independently invented by myself slightly later, under the name Annealed Importance Sampling — is a way of bypassing these difficulties.

- It uses a complicated importance sampling distribution, involving many MCMC updates (eg, Metropolis), pertaining to a sequence of distributions.
- We can't compute the density for this sampling distribution.
- **But:** We can use importance weights that don't require this density. Instead, the weights are products of density ratios involving intermediate states and intermediate distributions.

This works, but using these weights is likely to be less efficient than if we could use the true importance weights.

Properties of Hamiltonian Importance Sampling

I will describe a new importance sampling scheme, which can be used to estimate the partition function as well as equilibrium averages. This "Hamiltonian importance sampling" scheme has three desirable properties:

- It's exact, apart from round-off and statistical errors (no error from using a finite MD stepsize).
- It uses a annealing-style importance sampling distribution that will tend to visit various potential wells (eg, different conformations).
- We *can* compute the true importance weights for this importance sampling distribution.
- It cools the system by extracting energy (from the momentum) a bit at a time, so the system passes through all intermediate energy states.

The last property may be of pragmatic as well as theoretical importance, since it eliminates the need to determine a detailed schedule of temperatures for intermediate distributions (as in Jarsynski's method).

Probability Densities for Transformations of Variables

Before introducing the scheme, I'll review a crucial topic: How probabability densities transform.

Let the multi-dimensional variable x have density $\pi_x(x)$. Define a transformed variable $y = h(x)$, where h is differentiable and invertible.

The probability density for y is given by

$$\pi_y(y) \quad = \quad \pi_x(h^{-1}(y)) \,/\, |\det\, h'(h^{-1}(y))|$$

where $h'(x)$ is the Jacobian matrix for the transformation.

Simple example: If $y = \alpha x$, then $\pi_y(y) = \pi_x(y/\alpha)/\alpha^d$, where d is the dimensionality of x and y.

Basic Hamiltonian Importance Sampling

From now on, let's assume $x = (q, p)$ and $\pi(x)$ is proportional to $f(x) = \exp(-\beta H(q, p))$, with $H(q, p) = U(q) + p^T p/2$.

We define an importance sampling distribution for (q, p) as follows:

- Generate an initial value for q uniformly, and an initial value for p from its canonical distribution at some high temperature.

- Apply K leapfrog steps to move from this initial (q, p) to a final (q, p).
  Note: The Jacobian for each such transformation is one.

- After each leapfrog step, multiply p by some factor, α, less than one. This cools the system towards the desired lower temperature.
  Note: The Jacobian for this multiplication is α^d.

The randomness comes only from generation of the initial state. The Jacobian for the subsequent deterministic transformation is just α^{Kd}, so we can easily compute the density of the final point, and hence its importance weight.

Details of Basic Hamiltonian Importance Sampling

We generate each $x_i = (q_i, p_i)$ and associated weight, w_i, as follows:

1. Generate $q_i^{(0)}$ uniformly from its range (assumed bounded).
   Generate $p_i^{(0)}$ from its Gaussian canonical distribution at inverse temperature β_0, having density $K_0(p)$.

2. For $k = 1, \ldots, K$:

   Perform one (or more) leapfrog steps with stepsize ϵ to produce $(q_i^{(k)}, \tilde{p}_i^{(k)})$ from $(q_i^{(k-1)}, p_i^{(k-1)})$.

   Let $p_i^{(k)} = \alpha \tilde{p}_i^{(k)}$.

3. Let $q_i = q_i^{(K)}$ and $p_i = p_i^{(K)}$.

4. Let $w_i = \exp(-\beta H(q_i, p_i)) \;/\; (K_0(p_i^{(0)})/\alpha^{Kd})$, where d is the dimensionality of p (and q).

We will need to tune β_0, ϵ, α, and K to get good performance.

When Would We Expect This to Work?

For importance sampling to work well,

- All points typical of $\pi(x)$ must have a reasonably high probability of being sampled. **This is crucial.**
- Points not typical of $\pi(x)$ must not be sampled too often. But this is less crucial.

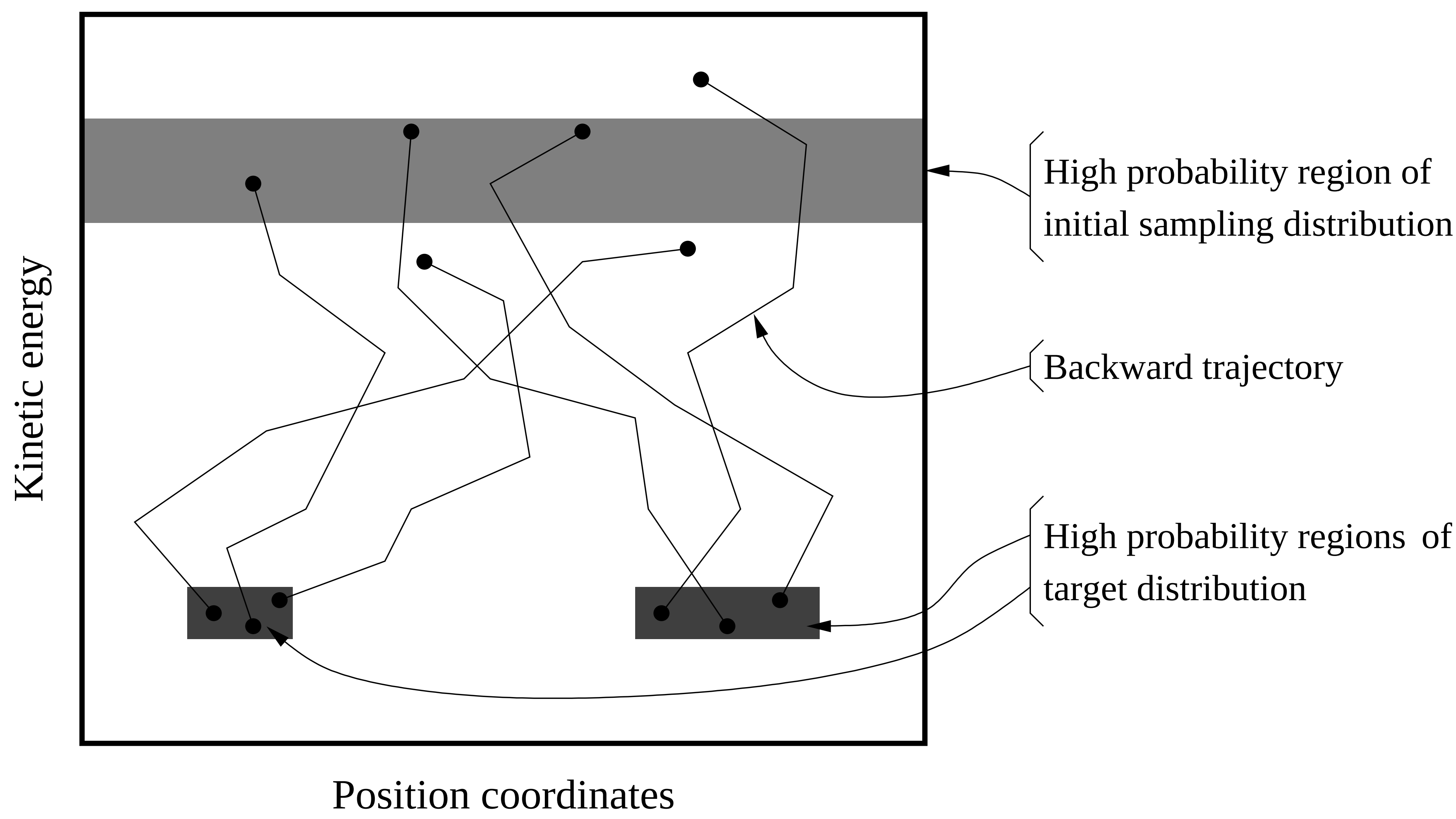
To check how well Hamiltonian Importance Sampling will work, we can imagine *backward* trajectories with *division* of p by α, starting from points drawn according to π. These backward trajectories must lead to points typical of the initial distribution (uniform for q, temperature $1/\beta_0$ for p).

There's reason to doubt this:

- We'd need to make a good guess at K to match the cooling time.
- There may be *no* good value for K, if there are multiple potential wells of different depths.

Picturing the Problem

Here's a picture of how the backward trajectories might not reach the region of high initial probability:



Picking the Number of Steps Randomly

We can fix this problem by choosing the number of leapfrog steps randomly, from some range, $K_{\rm min}, \ldots, K_{\rm max}$. If we choose K_i, we then use the same procedure as before to produce $(q_i, p_i) = (q_i^{(K_i)}, p_i^{(K_i)})$.

But: To compute the importance weight, we now need to add together the probability of generating (q_i, p_i) using *any* value for K, not just K_i.

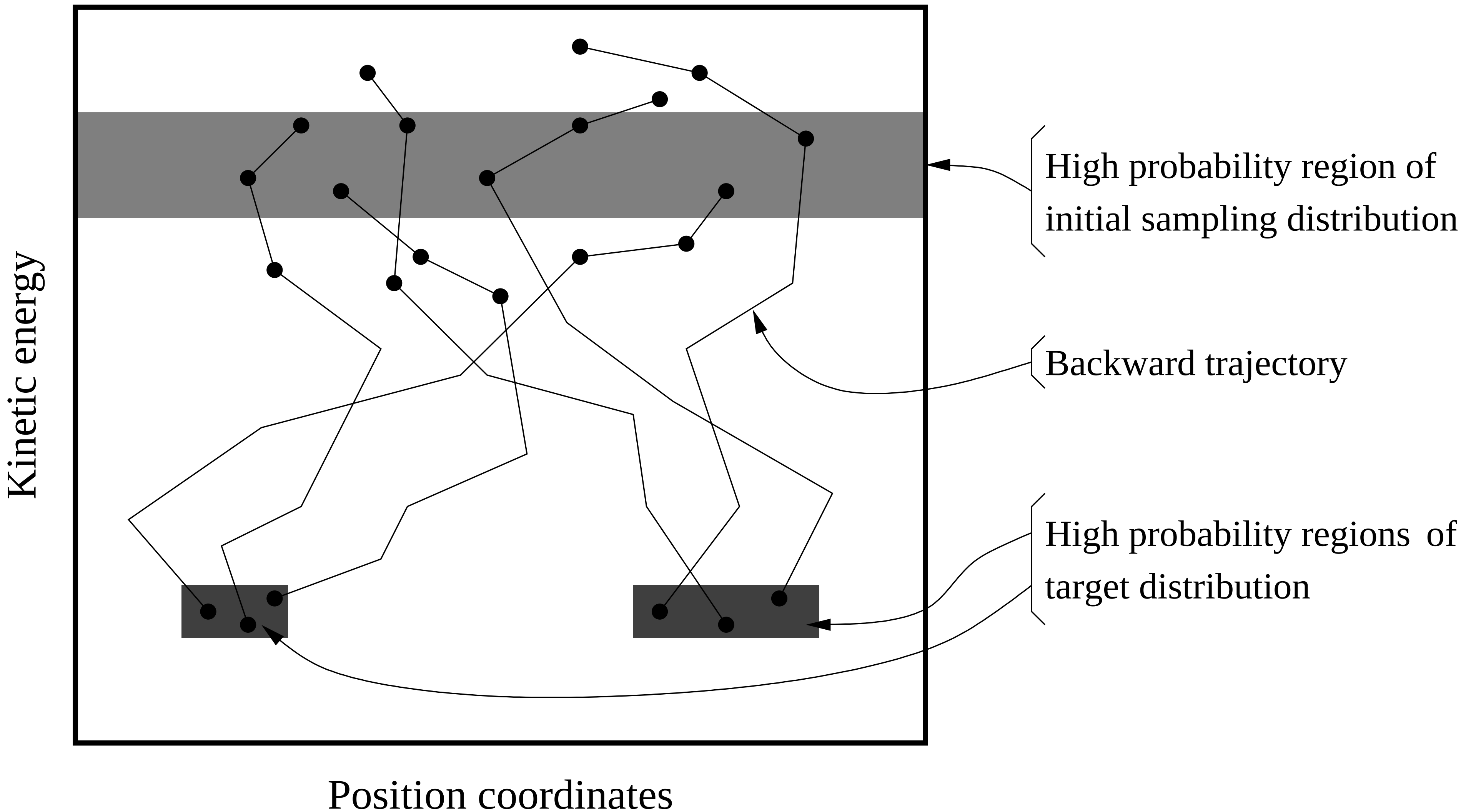
To do this, we simulate backwards (dividing p by α) from $(q_i^{(0)}, p_i^{(0)})$ for $K_{\rm max} - K_i$ leapfrog steps, to get $(q_i^{(-1)}, p_i^{(-1)}), \ldots, (q_i^{(K_i - K_{\rm max})}, p_i^{(K_i - K_{\rm max})})$.

The total probability of generating (q_i, p_i), ignoring the uniform density for q, can then be computed as

$$\frac{1}{K_{\rm max} - K_{\rm min} + 1} \sum_{K=K_{\rm min}}^{K_{\rm max}} K_0(p_i^{(K_i - K)}) / \alpha^{Kd}$$

Picturing this Solution

Here's how the problem seen before goes away if we randomizing the number of leapfrog steps to the previous number plus -1, 0, or $+1$:



Ensuring Equipartition of Kinetic Energy

Another potential problem: Backward trajectories from typical points may result in states at the initial temperature that aren't in equilibrium with respect to partition of kinetic energy among momentum variables.

Example: Backward trajectories from a cluster will lead to atoms escaping from the cluster at various times, with various kinetic energies, which may be unlikely to interact thereafter.

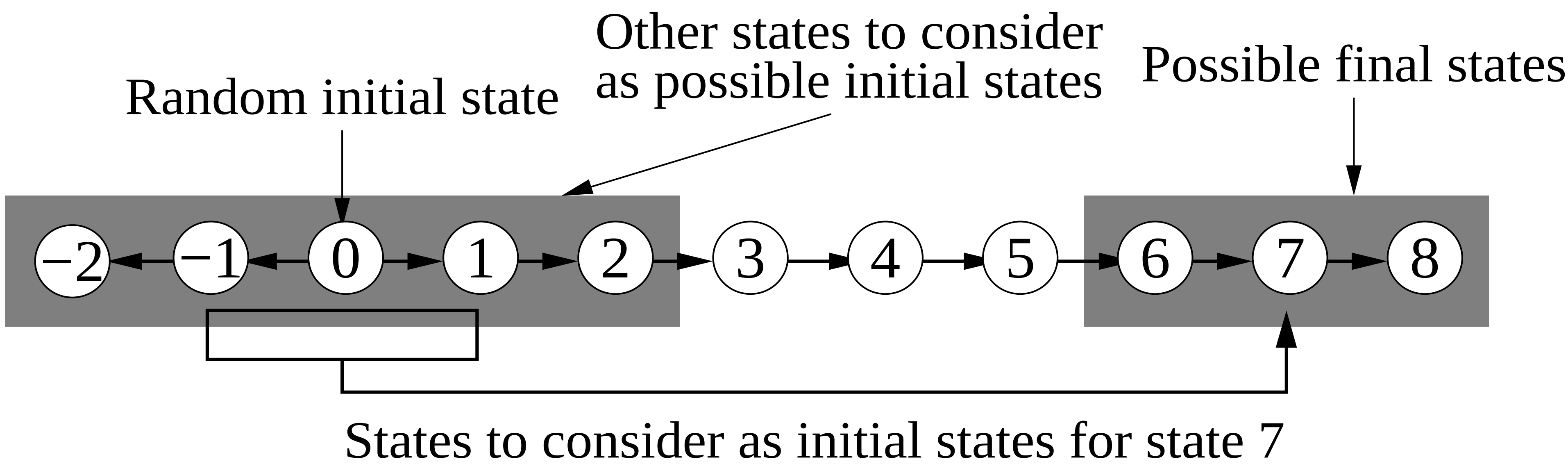
A solution: Periodically mix the momentum variables by doing a rotation in momentum space, using a series of random rotation axes and angles. Choosing randomly avoids the possibility that we're unlucky enough to fix on some particularly bad rotations, but for good performance, almost all the random choices must be good.

Simultaneously Producing Multiple Trajectories

Rather than get just one sampled state from a trajectory K_i steps long, with K_i randomly chosen from $K_{\min}$ to $K_{\max}$, we can with little extra effort get sampled states for *all* trajectory lengths from $K_{\min}$ to $K_{\max}$.

We just simulate forward for $K_{\max}$ steps, and backward for $K_{\max} - K_{\min}$ steps, then look at the $K_{\max} - K_{\min} + 1$ trajectories that start at the random initial state.

Here's a picture when $K_{\min} = 6$ and $K_{\max} = 8$:



Tests on 13-Atom Lennard-Jones Clusters

I tried Hamiltonian Importance Sampling on the simple problem of finding properties, including free energy, of 13-atom Lennard Jones clusters.

The atoms were in a 3D space with periodic boundary, with each dimension of length 10.

The LJ pair potential is

$$4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

I set $\epsilon = 1$ and $\sigma = 1$, and imposed an upper limit of 7.5 on the potential.

I looked at the canonical distribution at inverse temperature $\beta = 4$.

The initial distribution used was uniform for positions, and the canonical distribution at $\beta_0 = 1$ for momenta.

Results

I tried Hamiltonian Importance Samling with various settings of α, ϵ, $K_{\min}$, and $K_{\max}$.

Useful results were obtained using leapfrog steps with $\epsilon = 0.001$ (repeated 10 times), $\alpha = 0.9995$, $K_{\min} = 4000$, $K_{\max} = 7999$. With 500 trajectories, the result for free energy was $\log(Z_f/Z_g) \approx 57.87 \pm 0.32$, where Z_g is for an ideal gas at $\beta = 1$.

Better results were obtained (at five times the cost per trajectory) with $\epsilon = 0.001$ (repeated 5 times), $\alpha = 0.99995$, $K_{\min} = 40000$, $K_{\max} = 79999$. With 100 trajectories, the result was $\log(Z_f/Z_g) \approx 56.82 \pm 0.17$.

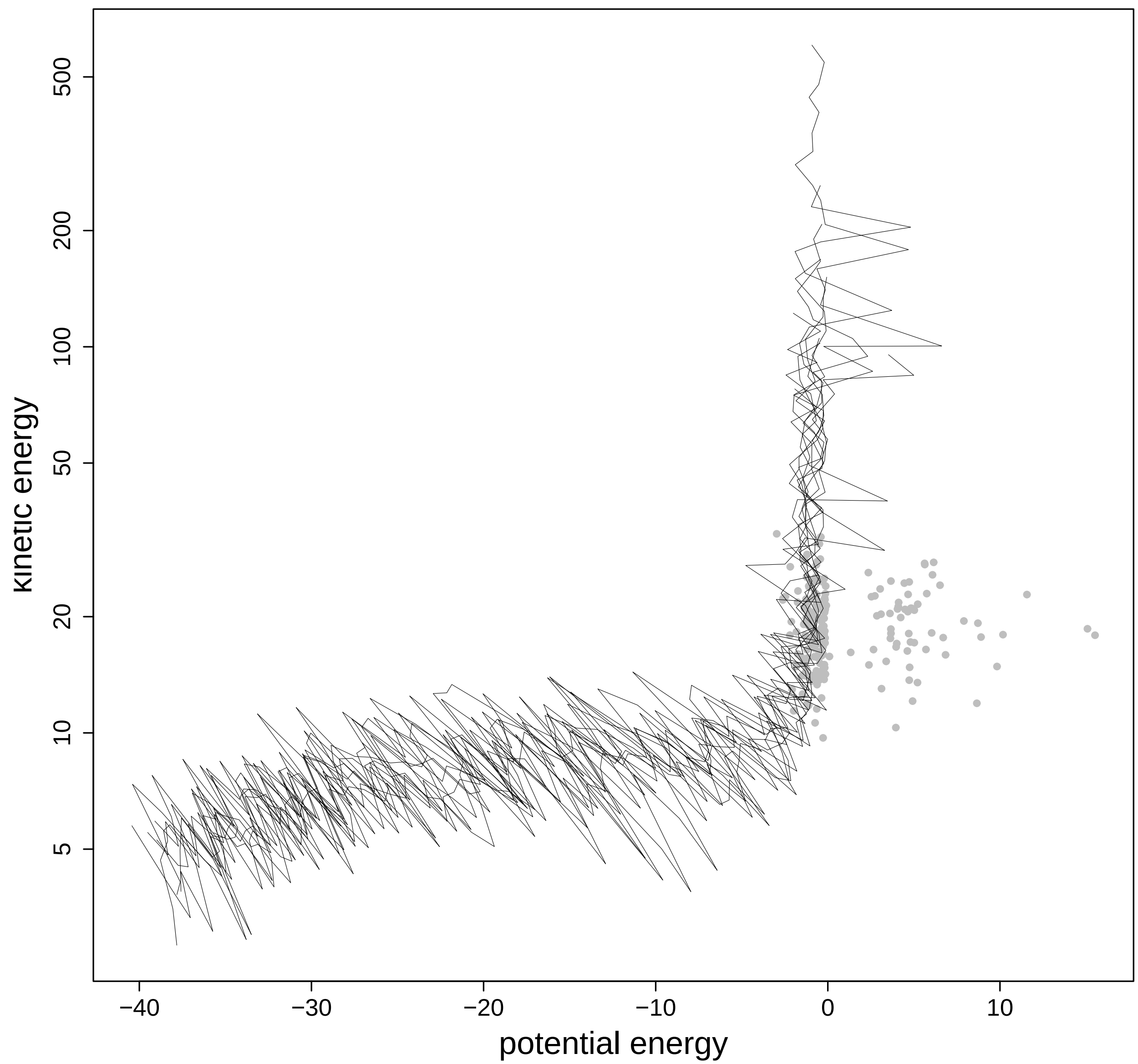
In both cases, momentum mixing to ensure equipartition was done. This turns out to be essential in this problem.

The result using Jarzynski's method, with 1000 runs using 4000 intermediate distributions, spaced manually to get good results, was $\log(Z_f/Z_g) \approx 56.90 \pm 0.11$.

All three of these methods took roughly the same amount of time.

A Test Using Backwards Trajectories

Let's check that we really are seeing the whole distribution by simulating backward trajectories from states gotten from a canonical MD simulation.

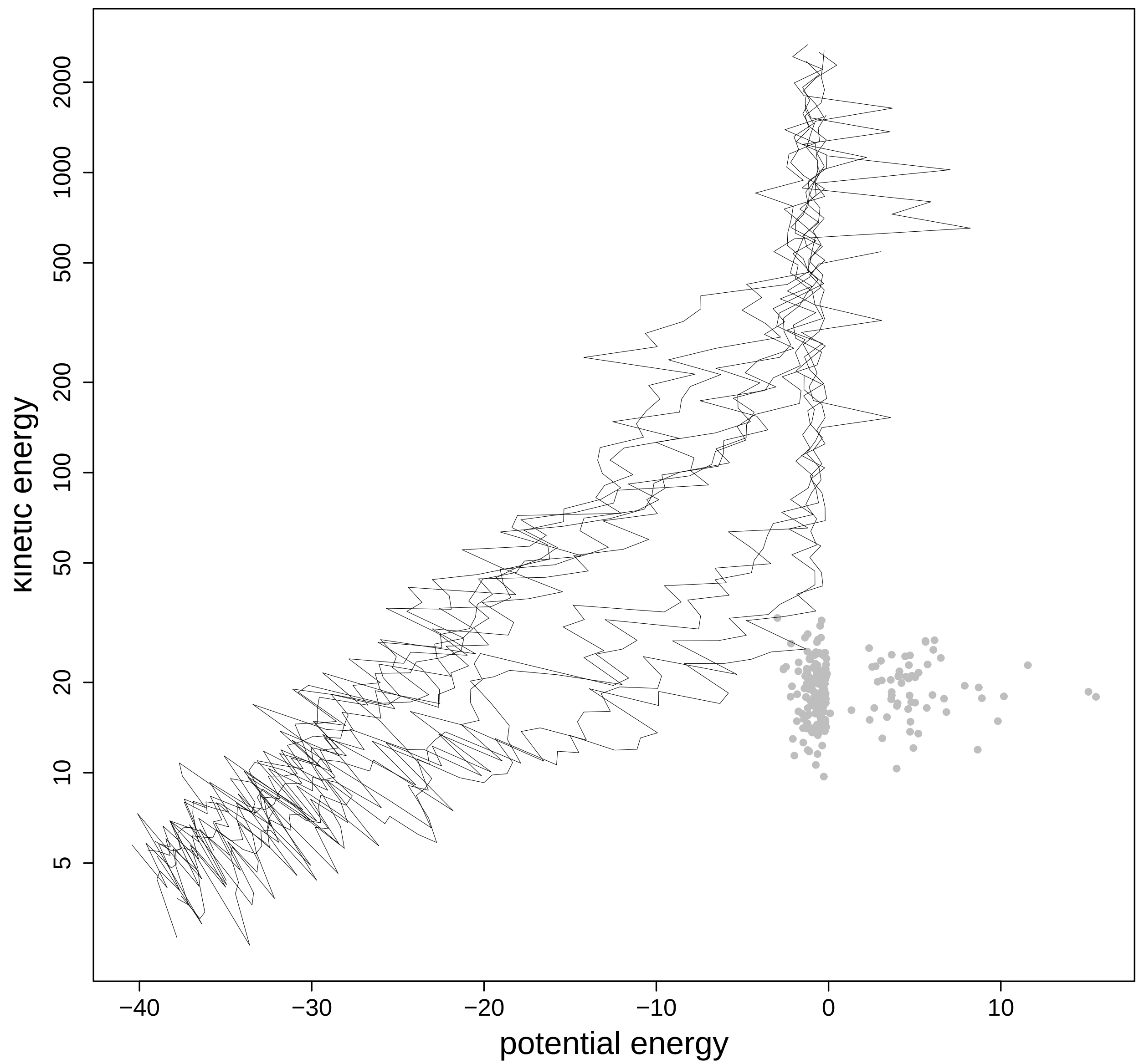


The gray dots are from the initial distribution (q uniform, p at $\beta = 1$). The lines are backward trajectories (first scheme from last slide), started from states from the canonical distribution at $\beta = 4$.

Note that all ten trajectories pass through a region of high initial probability.

Same Test But Without Momentum Mixing

If we omit the momentum mixing, equipartition is not maintained, and the method fails.



Now all but one of the trajectories misses the region of high initial probability.

To get good results without momentum mixing, one would need to use a much higher initial temperature.

Conclusions From the Tests

- Hamiltonian Importance Sampling can be applied successfully, at least to small problems. I've done preliminary work on larger problems with hundreds of atoms, in bulk, and I think this will work too. Using the NPT rather than NVT ensemble may help here in allowing a good initial distribution.

- Some time is "wasted" at present from using a small stepsize that may be needed only at the higher temperatures, and in simulating backward trajectories past the point where they could possibly matter.

- Efficiency is currently comparable to Jarzynski's method, but I hope that refinements will improve the comparison.

Future Work

- Refine the efficiency of the method — eg, figure out how to use variable stepsizes for varying temperatures.
- Try it out on various problems, including Bayesian statistical inference problems (my usual area of application).
- The same basic idea can be used in conjunction with Metropolis updates, with accept/reject decisions made deterministically based on how much energy is in a reservoir.
- Try to better understand the theory of such methods.
- Software implementing the method will be released soon. (This is only "toy" software, not meant for real MD applications go.)