Hamiltonian Importance Sampling

http://www.cs.utoronto.ca/~radford

Radford M. Neal, University of Toronto
We can estimate the ratio \( \frac{\mathcal{Z}'}{\mathcal{Z}} \) by \( \mathcal{Z}' / \int \mathcal{Z} \).

Here, \( \mathcal{Z}' \) is the importance weight for point \( x \).

Given points \( \{x_i\} \), drawn from \( \nu \), we can estimate the expectation of \( \nu \) with respect to \( \nu \), by

\[
\mathbb{E}_{\nu}[\ell(x)] = \frac{\int \nu(x) \ell(x) dx}{\int \nu(x) dx} = \frac{\mathcal{Z}' / \int \mathcal{Z}}{\int \nu(x) dx}.
\]

\( \mathcal{Z}' \) is the importance weight for point \( x \).

If \( \mathcal{Z}' \) is the importance weight for point \( x \),

\[
\mathcal{Z}' = \sum_{u} w_u \mathcal{Z}.
\]

\( w_u \) is the weight of point \( u \).

We want to estimate expectations with respect to the distribution with

\[ f(x) \int f(x) dx = \mathcal{Z} \]

We have

\[
\mathcal{Z} = \int f(x) dx.
\]

We can't sample from \( \nu \), so instead, we sample from the probability density proportional to \( \nu \),

\[
f(x) \int f(x) dx = \mathcal{Z}.
\]

We want to estimate expectations with respect to the distribution with

\[ f(x) = \frac{f(x)}{\int f(x) dx} \]

We have

\[
\mathcal{Z} = \int f(x) dx.
\]
involves an infeasible integral over all intermediate states.

the density for the distribution defined by \( Y \) Metropolis updates.

\[ \text{We can compute } g(x), \text{ and hence the importance weights. Sadly,} \]

(3) Starting from a uniformly-distributed start state.

something like the distribution defined by \( \bar{Y} \) Metropolis updates
distributions like Gaussians aren't good approximations. We need
dependently). Easy to sample from it.

(2) We can feasibly sample from it.

\[ f(x) \] will be very small.

highly variable, and the effective sample size when estimating \( \langle a \rangle \) not.

I) Is a good approximation to \( y \). If not, the importance weights will be

Difficulties with Importance Sampling
could use the true importance weights. This works, but using these weights is likely to be less efficient than if we intermediate states and intermediate distributions.

Instead, the weights are products of density ratios involving

\[ \text{But: We can use importance weights that don't require this density:} \]

\[ \text{We can't compute the density for this sampling distribution:} \]

\[ \text{We can't compute the density for this sampling distribution:} \]

\[ \text{It uses a complicated importance sampling distribution, involving} \]

\[ \text{these difficulties.} \]

under the name Annealed Importance Sampling — is a way of bypassing

Järzynski's method — independently invented by myself slightly later,
temperatures for intermediate distributions (as in Jarzynski’s method).

Since it eliminates the need to determine a detailed schedule of
The last property may be of pragmatic as well as theoretical importance,

at a time, so the system passes through all intermediate energy states.
It cools the system by extracting energy (from the moment) a bit

- sampling distribution.

We can compute the true importance weights for this importance

tend to visit various potential wells (i.e., different conformations).
It uses a simple-sampling-style importance sampling distribution that will

- using a finite MD stepsize.

It's exact, apart from round-off and statistical errors (no error from

"Hamiltonian Importance Sampling" scheme has three desirable properties.

This estimate the partition function as well as equilibrium averages. This

I will describe a new importance sampling scheme, which can be used to

Properties of Hamiltonian Importance Sampling
The probability density for $y$ is given by

$$|((f_i)_{1-\nu}, y)_{\nu} / ((f_i)_{1-\nu})_{\nu} = (f_i)^{\nu}_{\nu}$$

The probability density for $y$ is differentiable and invertible.

Let the multi-dimensional variable $x$ be density $x_{\nu}$. Define a

Let the multi-dimensional variable $y$ have density $y_{\nu}$. Define a

Before introducing the scheme, I'll review a crucial topic: How

Transformations of Variables for

Probabilistic Densities
Importance Weight.

so we can easily compute the density of the final point, and hence its
Jacobian for the subsequent deterministic transformation is just \( K \).

The randomness comes only from generation of the initial state. The

\[ p \cdot \frac{\partial}{\partial_t} \frac{\partial}{\partial x} \frac{\partial}{\partial y} \frac{\partial}{\partial z} \frac{\partial}{\partial w} \]

Note: The Jacobian for this multiplication is \( c^p \).

Note: The Jacobian for each such transformation is one.

\[ (d,b) \] (to a primal state) by some factor, \( c \), less than one.

After each leapfrog step, multiply by some factor, \( c \), less than one.

Note: The Jacobian for each such transformation is one.

Apply \( N \) leapfrog steps to move from this initial state to a primal state.

from its canonical distribution at some high temperature.

d for its uniform value for \( b \), uniformly, and an initial value for \( d \).

Generate an initial value for \( b \), uniformly, and an initial value for \( d \).

We define an importance sampling distribution for \((d, b)\) as follows:

\[ \frac{\pi}{d_L d} \neq \frac{1}{(d', b)H} \text{ with } \frac{\exp(\text{exponents})}{(x)}f \]

From now on, let's assume \( \frac{\pi}{d_L d} \) is proportional to \( (x)^\nu \) and \( (d', b) = x \).

Basic Hamiltonian Importance Sampling
We will need to tune $\theta_0$, $\epsilon$, $\alpha$, and $\lambda$ to get good performance. Dimensionality of $d$ and $b$ where $p$ is the $(r, \lambda) \alpha/(\lambda^2 d)^0 Y / ((\lambda^2 d, \lambda b) \theta \theta^{-}) \exp \equiv m \lambda e$.

1. Let $d$ = $\lambda b$ and $b$ = $\lambda b$.
2. Let $d_{(\lambda)} = d_{(\lambda)}$ from $(\lambda^2 d, \lambda b)$ to produce $(\lambda^2 d, \lambda b)$.

Perform one (or more) leapfrog steps with stepsize $e$.

2. For $\lambda = 1, \ldots, K$.

- temperature $\theta_0$ having density $\theta_0 Y$.

Generate from its Gaussian canonical distribution at inverse generate from its range (assumed bounded).

1. Generate uniformly from its range ($d^2 b$, $b$) = $b$ and associated weight, $w$ as follows:

Details of Basic Hamiltonian Importance Sampling
There are multiple potential wells of different depths. There may be no good value for $k$, if there are multiple potential wells, we'd need to make a good guess at $y$ to match the cooling time. There's a reason to doubt this.

Typical of the initial distribution (uniform for $q$ temperature $t$ for $p$), these backward trajectories must lead to points drawn according to $v$. These backward trajectories must lead to points by a starting from points imaged backward trajectories with division of $p$ by $x$. To check how well Hamiltonian Importance Sampling will work, we can check how well.

Points not typical of $v(x)$ must not be sampled too often. But this is crucial.

This is crucial.

For Importance Sampling to work well, we would expect this to work.
Here is a picture of how the backward trajectories might not reach the region of high initial probability.
\[ p_{\text{Y}} \sim \left( \frac{Y - \gamma}{\max \gamma - \gamma} \right)^{y_{\text{max}}} \left\{ \sum_{Y} \frac{1}{\gamma_{\text{max}} - \gamma} \right\} \]

For \( k \), can then be computed as the total probability of generating \( (d, b) \) ignoring the uniform density.

The procedure steps, to get \( Y - \gamma \) from \( d \) and \( b \) by dividing backwards (dividing by \( d \) or \( b \) for \( (d, b) \) from \( (0, 0) \)).

To do this, we simulate backwards, starting with \( \gamma \), not just \( Y \).

The probability of generating \( (d, b) \) using any pair \( (\gamma, \gamma) \) requires adding together

\[ \left( (\gamma, \gamma) \right) = (d, b) \]

use the same procedure as before to produce \( \gamma \).

If we choose \( Y \), we then randomly choose \( \gamma \).

We can fix this problem by choosing the number of steps randomly.

Pick the Number of Steps Randomly
Here’s how the problem seems before we randomizing the initial sampling distribution.

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

**Solution:** Periodically mix the momentum variables

which may be unlikely to interact thereafter.

**Example:** Backward trajectories from a cluster will lead to atoms escaping from the cluster at various times, with various kinetic energies,

**Problem:** Backward trajectories from typical points may result in states at the initial temperature that aren’t in equilibrium and new equilibrium partition of kinetic energy
Random initial state

Possible final states

States to consider as initial states for state 7

Possible initial states

Other states to consider as possible initial states

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

Random initial state.

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

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

Rather than get just one sampled state from a trajectory $K^*$ steps long, we can get $K^*$ randomly chosen from $Y_{\text{min}}$ to $Y_{\text{max}}$, we can with little extra effort get sampled states for all trajectory lengths from $Y_{\text{min}}$ to $Y_{\text{max}}$. Instead of producing multiple trajectories

Simultaneously Producing Multiple Trajectories
The initial distribution used was uniform for positions, and the canonical distribution at inverse temperature $\beta = \frac{1}{kT}$.

I looked at the canonical distribution at inverse temperature $\beta = \frac{1}{kT}$ on the potential.

$$[\left(\frac{\partial}{\partial \phi}\right)_{\phi=0} - \left(\frac{\partial}{\partial \phi}\right)_{\phi=\phi_0}]$$

The LJ pair potential is dimensionless.

The atoms were in a 3D space with periodic boundary conditions with each property, including free energy, of 10-atom Lennard-Jones clusters.

I tried Hamiltonian Importance Sampling on the simple problem of finding tests on 1-3 Atom Lennard-Jones Clusters.
All three of these methods took roughly the same amount of time.

\[ \log_8 \frac{6.80 \pm 0.11}{56.90 \pm 0.11} \approx \left( \frac{6Z}{fZ} \right) \]

intermediate distributions, spaced manually to get good results, was.

The result using Jarzynski's method, with 1000 runs using 4000

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

With 100 trajectories, the result was \( Y_{\text{mix}} = 0.99995 \), \( Y_{\text{mini}} = 0.99995 \), \( e = 0.001 \) (repeated 5 times), \( \alpha = 0.01 \) (repeated 5 times), \( Y_{\text{mix}} = 0.99995 \), \( Y_{\text{mini}} = 0.99995 \), \( e = 0.001 \) (repeated 5 times), \( \alpha = 0.01 \) (repeated 5 times).

Better results were obtained (at half times the cost per trajectory) with

\[ b \]

an ideal gas at \( T = 1 \).

For \( 6Z \) where is for

the result for the energy was \( Y_{\text{mix}} = 0.99995 \), \( Y_{\text{mini}} = 0.99995 \), \( e = 0.001 \) (repeated 5 times), \( \alpha = 0.01 \) (repeated 5 times).

Useful results were obtained using Leapfrog steps with \( e = 0.001 \) (repeated 5 times), \( \alpha = 0.01 \) (repeated 5 times).

Results
Note that all ten trajectories pass through a region of high initial probability. From the canonical distribution, the trajectories started from states (first scheme from last lines are backward trajectories). The form is $\dot{q} = \frac{\partial}{\partial q} \ln \mathcal{P}(q)$. The gray dots are from the backward trajectories from states gotten from a canonical MD simulation.

Let’s check that we really are seeing the whole distribution by simultaneously using backwards trajectories.

A Test Using Backwards Trajectories
higher initial temperature. To get good results with a much higher initial probability, vectors miss the region of the test. Now all but one of the trajectories miss the equi-

potential energy kinetic energy

potential energy kinetic energy

The method fails. If we omit the momentum mixing, equilibrium is not maintained, and...
that refinements will improve the comparison. Efficiency is currently comparable to Jarzynski’s method, but I hope backward trajectories past the point where they could possibly matter may be needed only at the higher temperatures, and in simultaneous time is “wasted” at present from using a small stepsize that allowing a good initial distribution. Using the NPT rather than NVT ensemble may help here in problems with hundreds of atoms, in bulk, and I think this will work least to small problems. I’ve done preliminary work on larger Hamiltonian Importance Sampling can be applied successfully, at

Conclusions From the Tests
only "toy software, not meant for real MD applications". 

Software implementing the method will be released soon. (This is •

Try to better understand the theory of such methods. •

on how much energy is in a reservoir. 
The same basic idea can be used in conjunction with Metropolis •

inference problems (my usual area of application). •

Try it out on various problems, including Bayesian statistical •

variable stepsizes for varying temperatures. •

Refine the efficiency of the method — e.g., figure out how to use •

Future Work