Markov Chain Sampling Using Hamiltonian Dynamics

Radford M. Neal

Dept. of Statistics and Dept. of Computer Science
University of Toronto

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

Joint Statistical Meetings, Baltimore, August 1999

The Metropolis-Hastings Algorithm

We can sample from a distribution with density $\pi(x)$ by simulating a Markov chain with transitions defined as follows:

From the current state, x, a candidate state, x^*, is drawn from a proposal distribution, $S(x, x^*)$. The proposed state is accepted as the next state of the Markov chain with probability

$$\min\left[1,\ \frac{\pi(x^*)S(x*, x)}{\pi(x)S(x, x*)}\right]$$

If x^* is not accepted, the next state is the same as the current state.

If the proposal distribution is symmetric — ie, $S(x, x^*) = S(x^*, x)$ — then the acceptance probability depends only on $\pi(x^*)/\pi(x)$.

Proposals for the Metropolis Algorithm

What proposal distribution, $S(x, x^*)$, should we use? Two possibilities:

Independent Proposals: $S(x, x^*)$ does not dependent on x. Every proposal is from a fixed distribution that we select to be close to π.

Advantage: We can move far in one step.

Disadvantage: The rejection rate will be high when we can't find a good approximation to π.

Random Walk Proposals: $S(x, x^*)$ depends only on $x^* - x$. Each proposal is from a distribution centred on the current state.

Advantage: We can get a high acceptance rate by not trying to move too far in one step.

Disadvantage: If we move in small steps, it takes a long time to move a long distance.

Can we find a way to propose states that are

far from the current state

and that will be

accepted with high probability?

Deterministic Proposals

The proposals I will discuss are not random: $x^* = s(x)$, for some deterministic function s.

What symmetry conditions on s are needed for validity of the Metropolis algorithm, using min$[1, \pi(x^*)/\pi(x)]$ as the acceptance probability?

Proposal functions must be reversible:
If $x^* = s(x)$, then $x = s(x^*)$.

Proposal functions must preserve volume:
Their Jacobian must have absolute value one.

Example of a valid proposal: $x^* = -x$.

Examples of invalid proposals:

$x^* = x + 1$	Not reversible
$x^* = 1/x$	Jacobian is not one

Of course, we will need to use other, random transitions too, to get an ergodic Markov chain.

Hamiltonian Dynamics

Proposal functions that are reversible and volume-preserving can be obtained using Hamiltonian dynamics.

Let the state consist of n "position" variables, q_i, and n "momentum" variables, p_i. This state changes through "time" according to

$$\frac{dq_i}{dt} = +\frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

where $H(q,p)$ is the "Hamiltonian" function.

Two properties of this dynamics:

- It preserves volume in (q,p) space.
- It keeps $H(q,p)$ constant.

We can simulate this dynamics approximately with some small time step. This can be done so volume is preserved exactly, but H won't stay exactly constant.

From Probability Density to Energy

We will start with a density $\pi(q) \propto \exp(-E(q))$ for the variables of interest.

We introduce p_i to be independent of the q_i, with independent standard normal distributions.

The joint probability density for q and p can be written as

$$\pi(q, p) \;\propto\; \exp(-H(q, p))$$

with the Hamiltonian function defined as

$$H(q, p) \;=\; E(q) \;+\; \sum_{i=1}^{n} p_i^2/2$$

$E(q)$ is called the "potential energy"; the second term is the "kinetic energy".

We hope to sample jointly for q and p, then just ignore the unneeded p variables.

The HMC Algorithm

Here is the "Hybrid Monte Carlo" algorithm of Duane, Kennedy, Pendleton, and Roweth (1987).

Alternately perform the following two steps:

1) Draw new values for the p_i independently from standard normal distributions. This is a Gibbs sampling update.

2) Perform a Metropolis update, with the candidate state found by simulating Hamiltonian dynamics for time T and then negating the p_i.

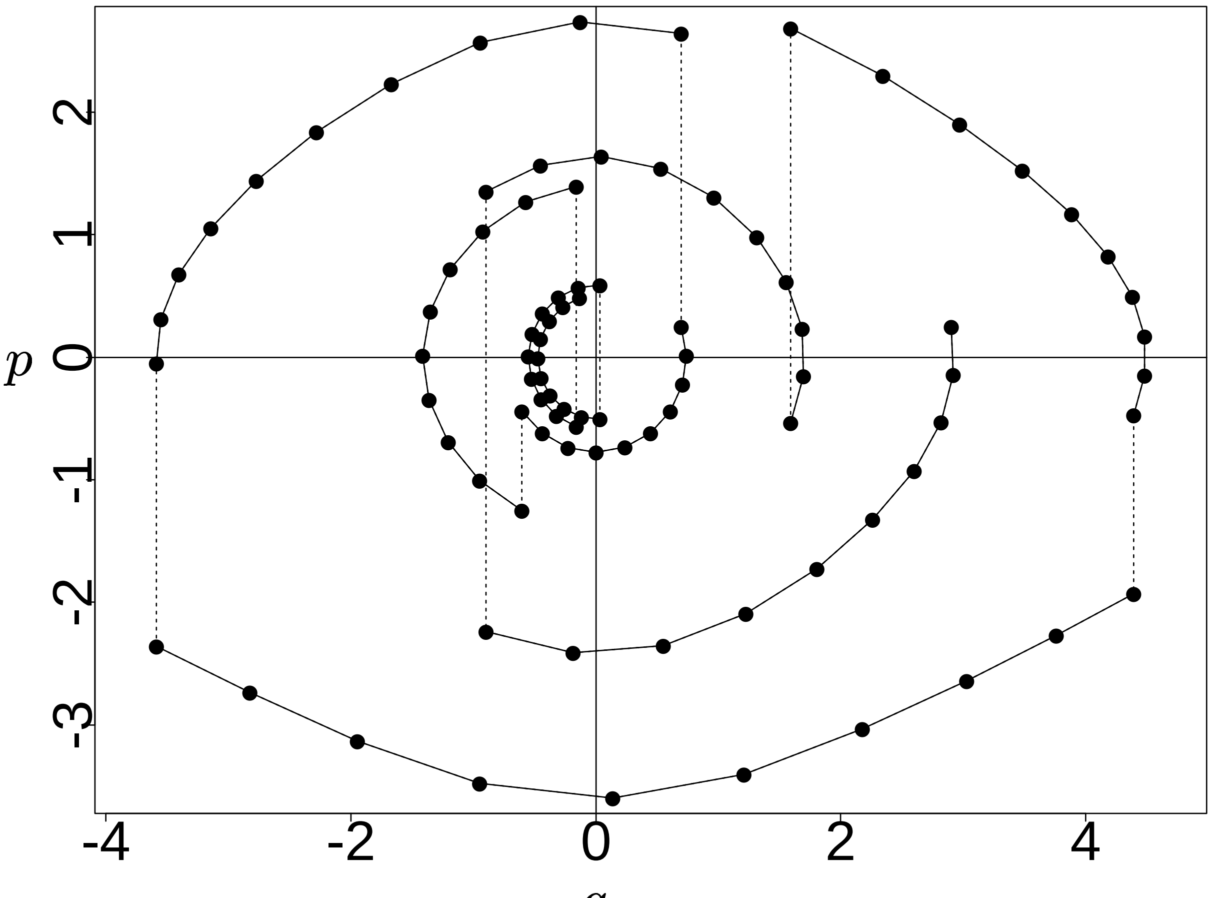
With $H(q,p) = E(q) + \sum p_i^2/2$, the dynamics is

$$\frac{dq_i}{dt} \;=\; p_i, \quad \frac{dp_i}{dt} \;=\; -\frac{\partial E}{\partial q_i}$$

It is easy to see that if we negate the p_i and then continue, we will retrace our path. The proposal function is therefore reversible, and since it also preserves volume, it is valid.

Example: Univariate t Distribution

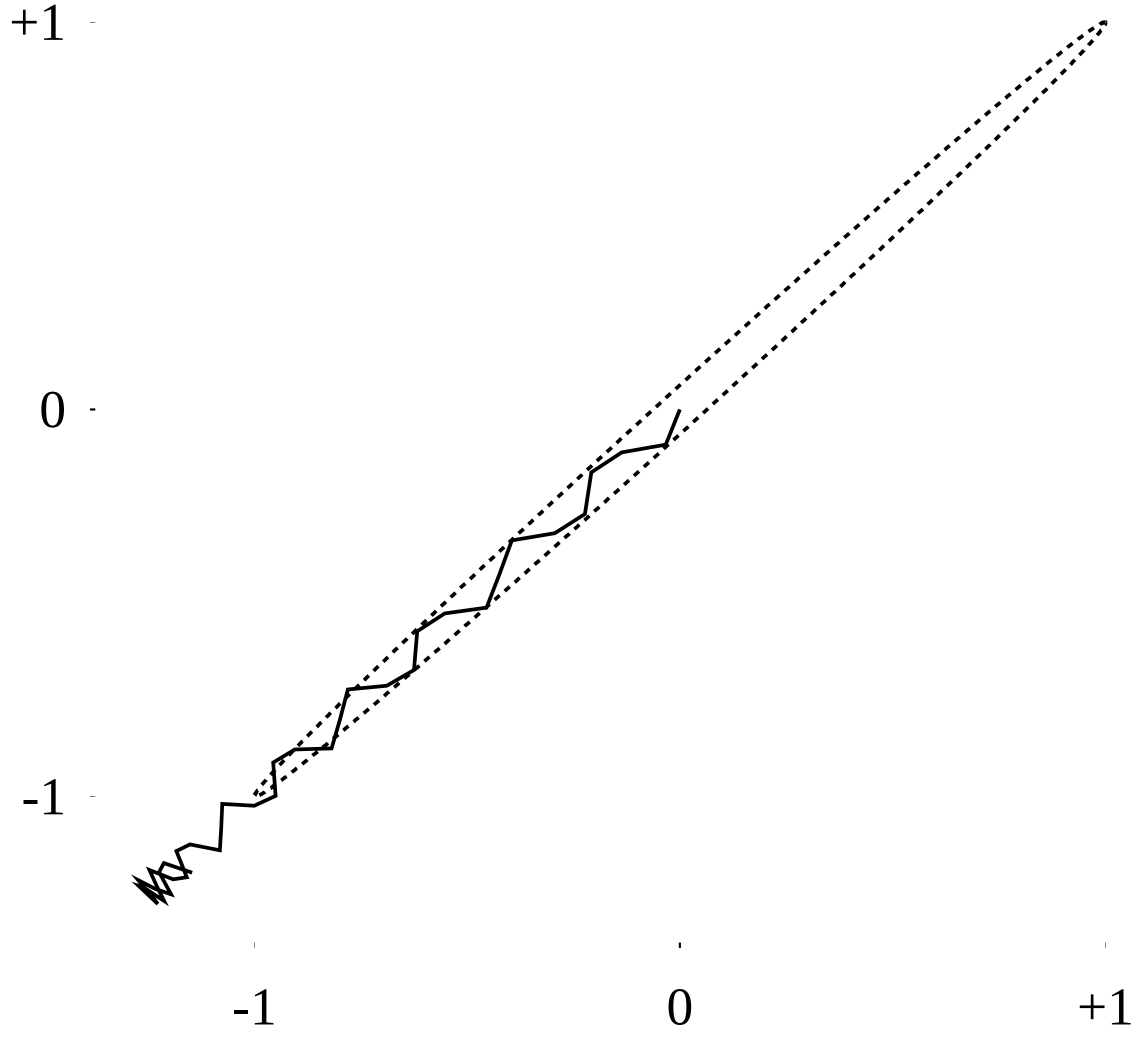
Here are some steps of HMC sampling from a t distribution with 5 degrees of freedom, for which $E(q) = 3\log(1 + q^2/5)$:



The time step used to simulate the dynamics above is smaller than needed to keep H almost constant, and hence accept most proposals.

Example: Bivariate Gaussian

Here is a trajectory for sampling a bivariate Gaussian with $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.998$:



The plot shows q_1 and q_2, but not the corresponding momentum variables, p_1 and p_2.

Advantages of HMC

HMC can propose and accept distant points. Computing such a proposal may require taking many small steps, but...

The steps HMC takes can move in one direction consistently. Hence in k steps, we can move a distance proportional to k, not $\sqrt{k}$, as for random walk Metropolis updates.

HMC can quickly change the probability density. With n dimensions, the momentum replacement changes the log density by order $\sqrt{n}$. A single Metropolis update changes the log density by order 1 (Caracciolo, *et al* 1994).

This is a reason why HMC may work better than other ways of proposing distant points. With n dimensions, at least order n transitions are needed to reach an independent point using only Metropolis updates.

But HMC Doesn't Solve Everything!

HMC can be trapped in isolated modes. Each point along a trajectory is distributed approximately according to π, so trajectories are unlikely to pass through the low-probability points that separate the modes.

This can be addressed by "tempering" the trajectories (Neal 1996).

HMC can be slow to move to points with vastly different probability density. Since H is almost conserved along a trajectory, E for the proposed point can differ by only order $\sqrt{n}$ (what the kinetic energy is likely to vary by).

This is not a problem for roughly Gaussian distributions. It is for high-dimensional skewed distributions, and when there are modes with different widths, but similar total probability.

Can we find a way to propose states that have

vastly different probability density

and that will be

accepted with high probability?

Sampling with Spirals

By abandoning preservation of volume, we can propose and accept states with vastly different probability density — since the acceptance criterion will involve not only $\pi(x^*)/\pi(x)$, but a Jacobian factor accounting for volume change.

Here is an update for the "spiral" method:

1) Choose k uniformly from $0 \ldots K$.

2) From the current state, (q, p), simulate k steps of Hamiltonian dynamics. Before and after each step, multiply all the p_i by $\sqrt{\alpha}$. Number these states from 1 to k.

3) From the original state, simulate $K - k$ reversed steps of Hamiltonian dynamics. Before and after each step, divide the p_i by $\sqrt{\alpha}$. Number the states from -1 to $k - K$.

4) Select the next state from among all $K + 1$ states seen above (0 is current state), with probabilities proportional to $\pi(q^{(i)}, p^{(i)})\,\alpha^i$.

The Double Spiral Method

The spiral method won't move easily between isolated modes — by the time it reaches the other mode, the momentum will be too large to stay there.

The "double spiral" method solves this.

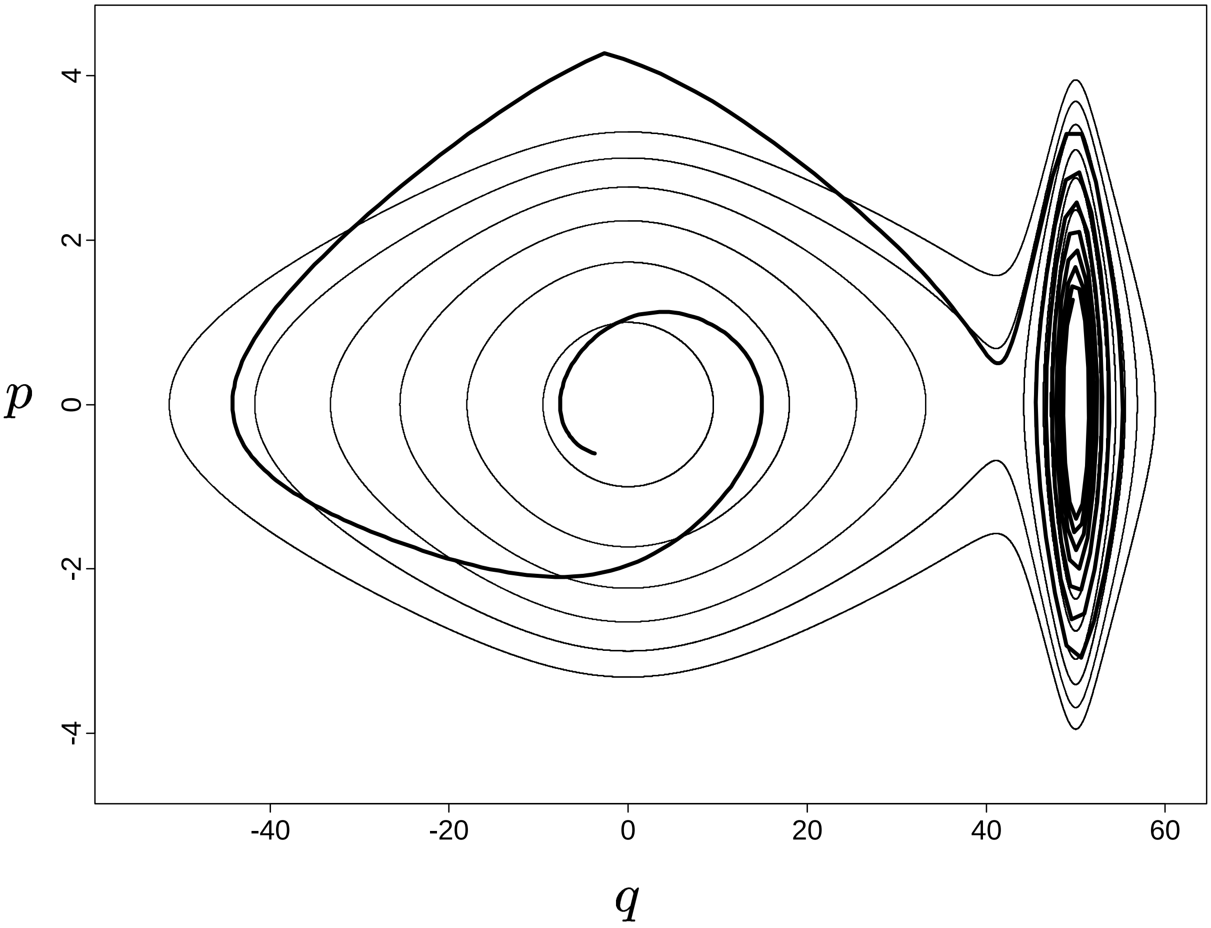
As for the spiral method, we simulate steps of Hamiltonian dynamics both ways from the current state, and multiply or divide by $\sqrt{\alpha}$ before and after each step.

But: We now switch from multiplying to dividing after a randomly selected number of steps. This switches us from an expanding spiral to a contracting spiral, which can converge on another mode.

Unlike my previous "tempered trajectory" method, there's a good chance of going to a mode with greatly different probability density.

Example: Bimodal Mixture

Here is a double spiral trajectory for an equal mixture of two t_5 distributions, centered at 0 and 50, with widths of 10 and 1:



If the start point was a typical point in one mode, there is a good chance that the next state will be a point in the other mode — even if it has much lower probability density.

References

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987) "Hybrid Monte Carlo", *Physics Letters B*, vol. 195, pp. 216-222.

Caracciolo, S., Pelissetto, A, and Sokal, A. D. (1994) "A general limitation on Monte Carlo algorithms of Metropolis type", *Physical Review Letters*, vol. 72, pp. 179-182.

Neal, R. M. (1993) *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 144 pages. Available from `http://www.cs.utoronto.ca/~radford/`

Neal, R. M. (1996) *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics No. 118, New York: Springer-Verlag.

Neal, R. M. (1994) "An improved acceptance procedure for the hybrid Monte Carlo algorithm", *Journal of Computational Physics*, vol. 111, pp. 194-203.

Neal, R. M. (1996) "Sampling from multimodal distributions using tempered transitions", *Statistics and Computing*, vol. 6, pp. 353-366.