12: Computationally Efficient Markov chain Simulation

11/18/19

HMC

Hamiltonian Monte Carlo can be quite effective at sampling from a high-dimensional posterior. It makes use of the derivative of the log-likelihood as well.

We will describe it in three steps:

1. Describing Hamiltonian dynamics in continuous time
2. Describing how to discretize Hamiltonian dynamics
3. Describing how to use these in a proposal distribution in the Metropolis-Hastings algorithm.

HMC

Say you have $\theta_1, \ldots, \theta_d$. You add d auxiliary variables: $\phi_1, \ldots, \phi_d$.

It's customary to use the notation $q_1, \ldots, q_d$ (the positions), and $p_1, \ldots, p_d$ (the momenta).

HMC

Say you have $\theta_1, \ldots, \theta_d$. You add d auxiliary variables: $\phi_1, \ldots, \phi_d$.

It's customary to use the notation $q_1, \ldots, q_d$ (the positions), and $p_1, \ldots, p_d$ (the momenta).

A HMC proposal follows a two-step procedure:

1. sample a random momentum vector
2. transform the momentum and position **nonrandomly** using Hamilton's equations

Both steps are transition kernels that preserve the stationary distribution.

HMC: a 1-d example

Start with one-dimensional q (position) and one-dimensional p (momentum). Also, m is mass (a tuning parameter).

1. potential energy: $U(q)$ is negative the logarithm of the unnormalized posterior.
2. kinetic energy: $K(p) = \frac{p^2}{2m}$

$p = m \times$ velocity

Two good resources:

1. `https://arxiv.org/pdf/1206.1901.pdf` (primary reference),
2. `https://arxiv.org/pdf/1701.02434.pdf`

HMC

When the particle goes up the hill, it loses kinetic energy, and gains potential energy.

Define the **Hamiltonian** as

$$H(q,p) = U(q) + K(p).$$

and define **Hamilton's Equations** as

$$\frac{dq}{dt} = \frac{\partial H(q,p)}{\partial p} \tag{1}$$

$$\frac{dp}{dt} = -\frac{\partial H(q,p)}{\partial q} \tag{2}$$

HMC: a first example

Assume the posterior is a Gaussian with mean $y = 0$ and variance 1. Negative log of the posterior is proportional to

$$U(q) = \frac{q^2}{2}.$$

Also, assume kinetic energy is of the form

$$K(p) = \frac{p^2}{2m}.$$

HMC: a first example

Assume the posterior is a Gaussian with mean $y = 0$ and variance 1. Negative log of the posterior is proportional to

$$U(q) = \frac{q^2}{2}.$$

Also, assume kinetic energy is of the form

$$K(p) = \frac{p^2}{2m}.$$

so

$$\frac{dq}{dt} = \frac{\partial H(q,p)}{\partial p} = \frac{dK(p)}{dp} = p(t)/m \tag{3}$$

$$\frac{dp}{dt} = -\frac{\partial H(q,p)}{\partial q} = -\frac{dU(q)}{dq} = -q(t) \tag{4}$$

HMC: a first example

If $m = 1$, a solution to

$$\frac{dq}{dt} = \frac{\partial H(q,p)}{\partial p} = \frac{dK(p)}{dp} = p(t) \tag{5}$$

$$\frac{dp}{dt} = -\frac{\partial H(q,p)}{\partial q} = -\frac{dU(q)}{dq} = -q(t) \tag{6}$$

is

$$q(t) = r\cos(a+t) \tag{7}$$

$$p(t) = -r\sin(a+t) \tag{8}$$

HMC: a first example

For this particular model, $(q, p)'$ rotates clockwise in phase-space.

$$q(t) = r\cos(a+t) \tag{9}$$

$$p(t) = -r\sin(a+t) \tag{10}$$

Can be written as

$$\begin{bmatrix} r\cos(a+t) \\ -r\sin(a+t) \end{bmatrix} = \underbrace{\begin{bmatrix} \cos([t-s]) & \sin([t-s]) \\ -\sin([t-s]) & \cos([t-s]) \end{bmatrix}}_{T_{t-s}} \begin{bmatrix} r\cos(a+s) \\ -r\sin(a+s) \end{bmatrix}$$

(use Angle-Sum trig identity)

HMC: a first example

When $m \neq 1$, T_s might look like this.

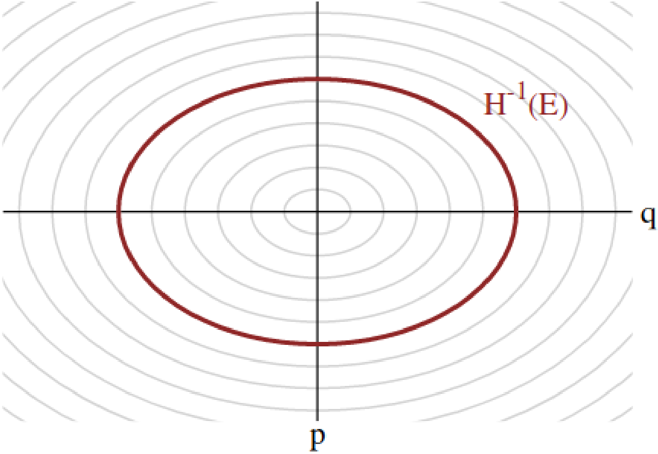


image source: `https://arxiv.org/pdf/1701.02434.pdf`

HMC

Now assume d-dimensional posterior $\mathbf{q} = (q_1, \ldots, q_d)$ and $\mathbf{p} = (p_1, \ldots, p_d)$.

When

$$K(\mathbf{p}) = \frac{\mathbf{p}' M^{-1} \mathbf{p}}{2}$$

Hamilton's equations become

$$\frac{dq_i}{dt} = \frac{\partial H(\mathbf{q}, \mathbf{p})}{\partial p_i} = \frac{dK(\mathbf{p})}{dp_i} = [M^{-1}\mathbf{p}]_i \tag{11}$$

$$\frac{dp_i}{dt} = -\frac{\partial H(\mathbf{q}, \mathbf{p})}{\partial q_i} = -\frac{dU(\mathbf{q})}{dq_i} \tag{12}$$

Recall that $U(\mathbf{q})$ is the negative log of the posterior you're interested in.

Property 1: Reversibility of T_s

$T_s : [\mathbf{q}(0), \mathbf{p}(0)] \mapsto [\mathbf{q}(s), \mathbf{p}(s)]$ always has an easy-to-find inverse.

Proof: just take the negative of the derivatives.

Property 2: Conservation of the Hamiltonian

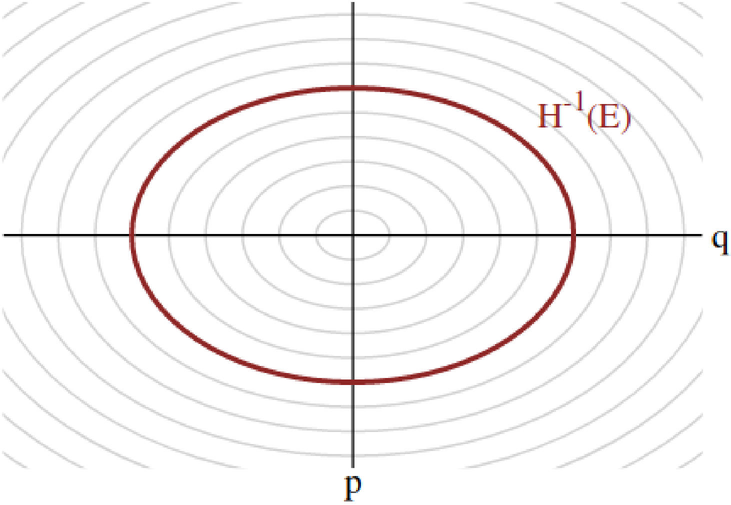
Using the chain rule:

$$\begin{aligned}
\frac{dH}{dt} &= \sum_{i=1}^{d} \frac{\partial H}{\partial p_i}\frac{dp_i}{dt} + \sum_{i=1}^{d} \frac{\partial H}{\partial q_i}\frac{dq_i}{dt} \\
&= \sum_{i=1}^{d} \frac{dK}{dp_i}\frac{dp_i}{dt} + \sum_{i=1}^{d} \frac{dU}{dq_i}\frac{dq_i}{dt} \\
&= \sum_{i=1}^{d} \frac{dK}{dp_i}\left(-\frac{dU}{dq_i}\right) + \sum_{i=1}^{d} \frac{dU}{dq_i}\frac{dK}{dp_i} \\
&= 0
\end{aligned}$$

Moving through time keeps you on the same contour or level-set in the phase space.

HMC

T_s keeps you on a level-set/contour:



HMC Property 3 and 4: Volume Preservation and Symplecticness

Volume preservation: $\{(q,p) : (q,p) \in A\}$ and $\{T_s(q,p) : (q,p) \in A\}$ have the same volume.

Symplecticness: a nice property of the Jacobian (matrix of time derivatives) of T_s.

Another thing: when we approximate these dynamics in our proposal distribution, these properties are preserved!

HMC: looking back at the big picture

HMC will work as follows: given that we are currently at position $\mathbf{q}(t)$, we are going to sample a momentum vector (which puts us on one of the level-sets), and then we are going to follow T_s for a deterministic amount of time (how much time is a tuning parameter we decide on).

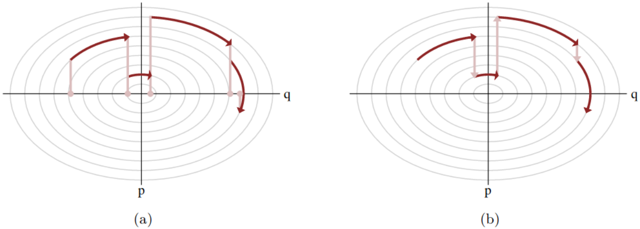


Fig 22. *(a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with a Hamiltonian trajectory before projecting back down to the target parameter space. (b) If we consider the projection and random lift steps as a single momentum resampling step, then the Hamiltonian Markov chain alternates between deterministic trajectories along these level sets (dark red) and a random walk across the level sets (light red).*

HMC: looking back at the big picture

HMC will work as follows: given that we are currently at position $\mathbf{q}(t)$, we are going to sample a momentum vector (which puts us on one of the level-sets), and then we are going to follow T_s for a deterministic amount of time (how much time is a tuning parameter we decide on).

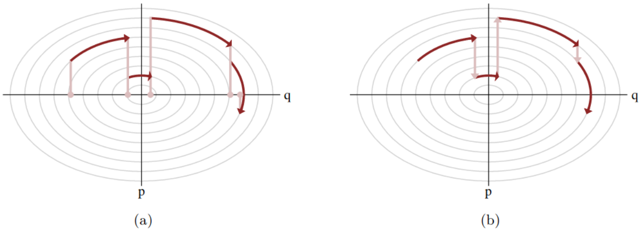


Fig 22. *(a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with a Hamiltonian trajectory before projecting back down to the target parameter space. (b) If we consider the projection and random lift steps as a single momentum resampling step, then the Hamiltonian Markov chain alternates between deterministic trajectories along these level sets (dark red) and a random walk across the level sets (light red).*

Following a contour line is impossible in continuous time though...

Discretizing Hamilton's Equations: Version 1.0

We need to be able to approximate T_s using the derivatives. To do that, we pick a small change in time called ϵ. Then we take L steps of size ϵ.

Two procedures are described. The last one is the one that is most commonly used.

For simplicity, assume the mass matrix is diagonal, making

$$K(\mathbf{p}) = \mathbf{p}' M^{-1} \mathbf{p} = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}.$$

Discretizing Hamilton's Equations: Version 1.0

When $K(\mathbf{p}) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$, **Euler's method** approximates the solution of

$$\frac{dq_i}{dt} = \frac{\partial H(\mathbf{q},\mathbf{p})}{\partial p_i} = \frac{dK(\mathbf{p})}{dp_i} = \frac{p_i}{m_i} \tag{13}$$

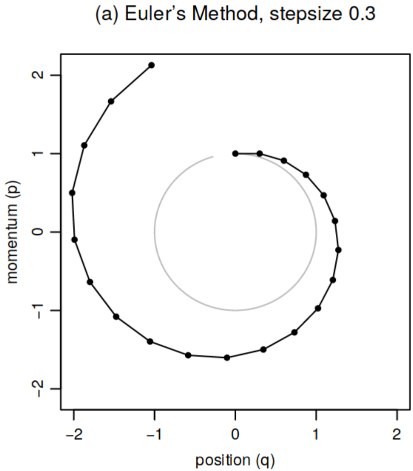
$$\frac{dp_i}{dt} = -\frac{\partial H(\mathbf{q},\mathbf{p})}{\partial q_i} = -\frac{dU(\mathbf{q})}{dq_i} \tag{14}$$

as

$$q_i(t+\epsilon) = q_i(t) + \epsilon \frac{p_i(t)}{m_i} \tag{15}$$

$$p_i(t+\epsilon) = p_i(t) - \epsilon \frac{dU(\mathbf{q}(t))}{dq_i} \tag{16}$$

Discretizing Hamilton's Equations: Version 1.0



Twenty steps when $H(q,p) = p^2/2 + q^2/2$, the initial state is $(q,p) = (0,1)$.

Discretizing Hamilton's Equations: Version 2.0

When $K(\mathbf{p}) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$, **the leap-frog method** approximates

$$\frac{dq_i}{dt} = \frac{\partial H(\mathbf{q},\mathbf{p})}{\partial p_i} = \frac{dK(\mathbf{p})}{dp_i} = p_i/m_i \tag{17}$$

$$\frac{dp_i}{dt} = -\frac{\partial H(\mathbf{q},\mathbf{p})}{\partial q_i} = -\frac{dU(\mathbf{q})}{dq_i} \tag{18}$$

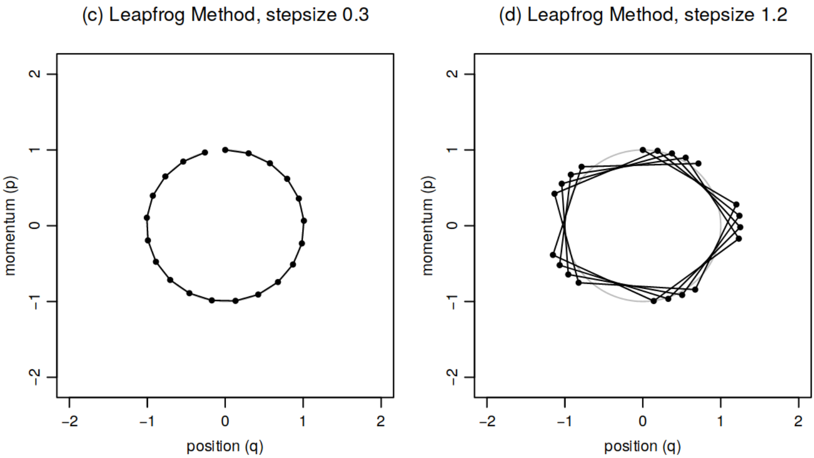
with

$$p_i(t+\epsilon/2) = p_i(t) - (\epsilon/2)\frac{dU(\mathbf{q}(t))}{dq_i} \tag{19}$$

$$q_i(t+\epsilon) = q_i(t) + \epsilon\frac{p_i(t+\epsilon/2)}{m_i} \tag{20}$$

$$p_i(t+\epsilon) = p_i(t+\epsilon/2) - (\epsilon/2)\frac{dU(\mathbf{q}(t+\epsilon))}{dq_i} \tag{21}$$

Discretizing Hamilton's Equations: Version 2.0


(c) Leapfrog Method, stepsize 0.3

(d) Leapfrog Method, stepsize 1.2

Twenty steps when $H(q,p) = p^2/2 + q^2/2$, the initial state is $(q,p) = (0,1)$.

Describing the HMC algorithm

The algorithm targets the distribution for $(\mathbf{q}, \mathbf{p})$:

$$\begin{aligned}
\frac{1}{Z} \exp\left[-\frac{H(\mathbf{q}, \mathbf{p})}{T}\right] &= \frac{1}{Z} \exp\left[-\frac{K(\mathbf{p}) + U(\mathbf{q})}{T}\right] \\
&= \frac{1}{Z} \exp\left[-\frac{K(\mathbf{p})}{T}\right] \exp\left[-\frac{U(\mathbf{q})}{T}\right] \\
&= \frac{1}{Z} \exp\left[-\frac{K(\mathbf{p})}{T}\right] \times \\
&\qquad \exp\left[-\frac{-\log\{\text{prior}(\mathbf{q}) \times \text{likelihood}(\mathbf{q})\}}{T}\right]
\end{aligned}$$

Describing the HMC algorithm

Step 1:

Sample p from the conditional target distribution

$$\frac{1}{Z} \exp\left[-\frac{K(\mathbf{p})}{T}\right].$$

In our case, this is the same as the marginal, due to independence.

Notice how this is a Gibbs-like step! It preserves the stationary distribution, and it has 100% chance of being accepted.

Describing the HMC algorithm

Step 2:

If we could integrate Hamilton's equations exactly, then our proposal would be deterministic, and we would accept with probability 1, because the Hamiltonian is preserved (property 2).

However, because we are using numerical leap-frog integration, there will be some change in the Hamiltonian. We think of the L leap-frog steps as a proposal distribution. This is a deterministic proposal, and it's symmetrical (we don't prove this). So what we end up with is a Metropolis-like acceptance probability:

$$\min\left[1, \frac{\exp\left[-H(\mathbf{q}^*, \mathbf{p}^*)\right]}{\exp\left[-H(\mathbf{q}, \mathbf{p})\right]}\right]$$

Describing the HMC algorithm

Here's a visualization:
`https://chi-feng.github.io/mcmc-demo/`