MCMC

Metropolis-Hastings and HMC

Chris Sherlock

Department of Mathematics and Statistics
Lancaster University

Outline

Bayesian Statistics

We have a data vector y, which is related to some parameters θ via a statistical model, giving $f(y|\theta)$ as the density (or mass function) of the data vector. It is also called the likelihood of θ.

We have some prior belief about the parameters (which we will assume are continuous) represented by a prior density $\pi_0(\theta)$.

Combining the prior and the likelihood using Bayes Theorem gives the posterior distribution for θ:

$$\pi(\theta) \equiv \pi(\theta|y) \propto \pi_0(\theta) f(y|\theta).$$

The constant of proportionality, $1/\int \pi_0(\theta) f(y|\theta)\, \mathrm{d}\theta$, is generally unknown.

Monte Carlo Estimation

We would like to evaluate, for example,

$$\begin{aligned} \mathbb{E}\left[\theta_1\right] & & , \\ \mathrm{Var}\left[\theta_1\right] &= \mathbb{E}\left[\theta_1^2\right] - \mathbb{E}\left[\theta_1\right]^2 , \\ \mathbb{P}\left(\theta_1 > 0\right) &= \mathbb{E}\left[\mathbb{I}_{\theta_1 > 0}\right] . \end{aligned}$$

All these quantities can be written in terms of expectations.

Let $h_n := \frac{1}{n}\sum_{i=1}^{n} h(\theta^{(i)})$, where $\theta^{(1)}, \ldots, \theta^{(n)}$ are samples from π.

The Strong Law of Large Numbers (SLLN) states that provided $\mathbb{E}_\pi\left[|h(\theta)|\right] < \infty$ then

$$h_n \rightarrow \mathbb{E}_\pi\left[h(\theta)\right].$$

So if we take a large enough sample (n) then we can estimate $\mathbb{E}\left[h(\theta)\right]$ as accurately as we wish.
In practice it is usually impossible to simply sample from π.

Markov chain Monte Carlo

A Markov Chain is a stochastic (random) sequence of values (or vectors) $\theta^{(1)}, \theta^{(2)}, \ldots$ such that

$$\mathbb{P}\left(\theta^{(t+1)}|\theta^{(1)}, \theta^{(2)}, \ldots \theta^{(t)}\right) = \mathbb{P}\left(\theta^{(t+1)}|\theta^{(t)}\right).$$

A Markov chain has a stationary density, $\pi(\theta)$, if

$$\theta^{(t)} \sim \pi \Rightarrow \theta^{(t+1)} \sim \pi,$$

which then implies that $\theta^{(T)} \sim \pi$ for all $T > t$.

An SLLN for Markov chains states that if a Markov chain has a proper stationary density, π, and provided $\mathbb{E}_\pi\left[|h(\theta)|\right] < \infty$ and the Markov chain is irreducible then

$$h_n := \frac{1}{n}\sum_{i=1}^{n} h(\theta^{(i)}) \to \mathbb{E}_\pi\left[h(\theta)\right].$$

So we need to construct a Markov chain with stationary density π.

Detailed balance

A Markov chain is said to satisfy detailed balance (DB) with respect to some density, π, if

$$\mathbb{P}\left(\theta^{(t)} = \theta, \theta^{(t+1)} = \theta'\right) = \mathbb{P}\left(\theta^{(t)} = \theta', \theta^{(t+1)} = \theta\right),$$

when the marginal density at time t is π.

$$\text{i.e. } \pi(\theta)P(\theta, \theta') = \pi(\theta')P(\theta', \theta),$$

where here $P(\theta, \theta')$ is the conditional density of the next value given the current value: $P(\theta, \theta') \equiv f_{\theta^{(t+1)}|\theta^{(t)}}(\theta'|\theta)$.

If a Markov chain satisfies DB with respect to π then the marginal density of the next value in the chain given that the current $\sim \pi$ is

$$\int \pi(\theta)P(\theta, \theta')\mathrm{d}\theta = \int \pi(\theta')P(\theta', \theta)\mathrm{d}\theta = \pi(\theta')\int P(\theta', \theta)\mathrm{d}\theta = \pi(\theta'),$$

so π is a stationary distribution of the Markov chain.

The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is

MH algorithm

Given current value $\theta^{(t)} = \theta$:
Propose a new value, θ' from some density $q(\theta'|\theta)$.
Define

$$\alpha(\theta, \theta') := 1 \wedge \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}.$$

With probability $\alpha(\theta, \theta')$ set $\theta^{(t+1)} = \theta'$ (accept), otherwise set $\theta^{(t+1)} = \theta$ (reject).

MH satisfies DB

The MH algorithm satisfies detailed balance with respect to π:

If $\theta' = \theta$ then the relationship is trivial since $\pi(\theta)P(\theta, \theta')$ and $\pi(\theta')P(\theta', \theta)$ both equal $\pi(\theta)P(\theta, \theta)$.

If $\theta' \neq \theta$ then there must be an acceptance so

$$\begin{aligned} \pi(\theta)P(\theta, \theta') &= \pi(\theta)q(\theta'|\theta)\alpha(\theta, \theta') \\ &= \pi(\theta)q(\theta'|\theta)\left(1 \wedge \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right) \\ &= \pi(\theta)q(\theta'|\theta) \wedge \pi(\theta')q(\theta|\theta'). \end{aligned}$$

This function does not change if θ and θ' are swapped, so it must equal $\pi(\theta')P(\theta', \theta)$.

Summary

Given a posterior density $\pi(\theta)$ we can construct a Markov chain which is irreducible and has π as its stationary density.

By the SLLN for Markov chains we can therefore estimate quantities such as

$$\mathbb{E}_\pi\left[\theta\right],\ \mathsf{Var}_\pi\left[\theta\right]\ \ \textit{or}\ \mathbb{P}_\pi\left(\theta > 0\right)$$

from the constructed Markov chain.

Of course one must choose a sensible $q(\theta'|\theta)$!

In what follows we will use x for the parameter vector of length $d \geq 1$, rather than θ.

HMC in a nutshell (in 1D)

The user chooses a value for a tuning parameter T.

Imagine a surface $U(x) = -\log \pi(x)$; our current position x is the position of a ball with mass m.

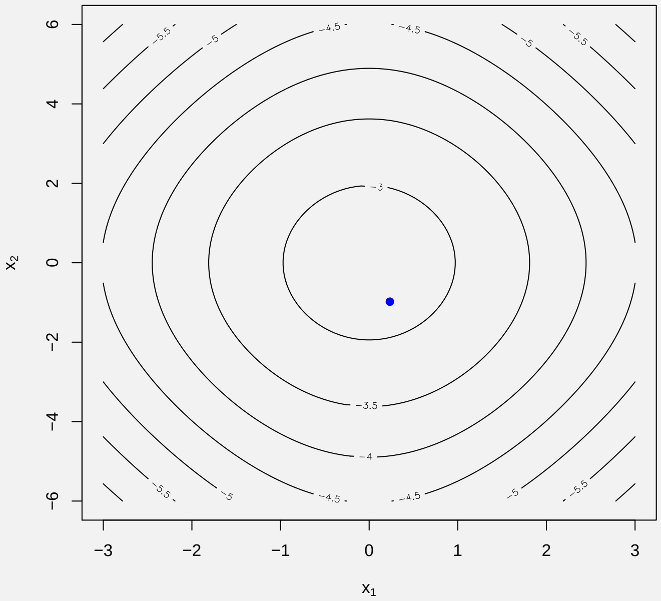
1. Kick the ball with a random amount of umpff in a random direction; i.e. give it a random momentum, p.
2. Watch it move along the surface for T seconds.
3. New position, x', is proposal for next point in Markov chain.
4. Accept proposal with a probability of

   $$\alpha = \min\left(1, \exp\left\{U(x) + \frac{p^2}{2m} - \left[U(x') + \frac{(p')^2}{2m}\right]\right\}\right),$$

   where p' is momentum at time T (formula explained later); else stay at x.
5. Go to 1.

Conservation of energy $\Rightarrow$ if we could integrate the dynamics exactly, $\alpha = 1$.

Potential and basics



How do we describe the movement in this potential?

Momentum is $p = mv$, where m is the mass and $v = \mathrm{d}x/\mathrm{d}t$ is the horizontal velocity.

Hamilton's Equations

$$\begin{aligned} \text{Potential energy:} \quad & U(x) \\ \text{Kinetic energy:} \quad & T(x) = \frac{1}{2}mv^2 = \frac{p^2}{2m} \\ \text{Hamiltonian:} \quad & H(x) = U(x) + T(x) \end{aligned}$$

Energy is conserved so

$$\begin{aligned} 0 = \frac{\mathrm{d}H}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}U(x) + \frac{p}{m}\frac{\mathrm{d}}{\mathrm{d}t}p \\ &= \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}x}U(x) + v\frac{\mathrm{d}}{\mathrm{d}t}p \end{aligned}$$

Hence

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\mathrm{d}U}{\mathrm{d}x}.$$

Also $$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{p}{m},$$

since $v = \mathrm{d}x/\mathrm{d}t$. *Show first animation.*

Solving Hamilton's Equations (1)

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\mathrm{d}U}{\mathrm{d}x}, \quad \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{p}{m}.$$

Intractable except in special cases. Hence, given (x_0, p_0), must solve numerically to obtain (x_T, p_T). Pick a time step, $\Delta t = \epsilon$.

Scheme 1 (Euler):

$$-\frac{\mathrm{d}U}{\mathrm{d}x} = \frac{\mathrm{d}p}{\mathrm{d}t} \approx \frac{\Delta p}{\Delta t} = \frac{p_\epsilon - p_0}{\epsilon},$$

so set

$$p_\epsilon = p_0 - \epsilon U'(x_0).$$

Similarly

$$x_\epsilon = x_0 + \frac{\epsilon}{m} p_\epsilon.$$

Repeat to obtain $(x_{2\epsilon}, p_{2\epsilon}), \ldots, (x_T, p_T)$. Error in $(x_T, p_T) \propto \epsilon$.

Solving Hamilton's Equations (2)

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\mathrm{d}u}{\mathrm{d}x}, \quad \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{p}{m}.$$

Given (x_0, p_0), we require (x_T, p_T). Pick a time step, $\Delta t = \epsilon$.

Scheme 2 (Leapfrog):

$$p_{\epsilon/2} = p_0 - \frac{\epsilon}{2} U'(x_0),$$

$$x_\epsilon = x_0 + \frac{\epsilon}{m} p_{\epsilon/2},$$

$$p_\epsilon = p_{\epsilon/2} - \frac{\epsilon}{2} U'(x_\epsilon).$$

Error in $(x_T, p_T) \propto \epsilon^2$- and much, much more! *Show 2nd and 3rd animations.*

Leapfrog is skew-symmetric

If, starting from (x_0, p_0) and integrating for one timestep using the leapfrog scheme, we obtain (x_ϵ, p_ϵ) then starting from $(x_\epsilon, -p_\epsilon)$ we would obtain $(x_0, -p_0)$ (look again at scheme).

Hence if, starting from (x_0, p_0) we eventually obtain (x_T, p_T) then starting from $(x_T, -p_T)$ we would eventually obtain $(x_0, -p_0)$.

Leapfrog has a Jacobian of 1

The leapfrog is a transformation $L : (x_0, p_0) \to (x_\epsilon, p_\epsilon)$. So, the density of (X_ϵ, P_ϵ) satisfies

$$f_{X_\epsilon, P_\epsilon}(x_\epsilon(x_0, p_0), p_\epsilon(x_0, p_0))|\det J| = f_{X_0, P_0}(x_0, p_0),$$

where

$$J = \left[\begin{array}{cc} \frac{\partial x_\epsilon}{\partial x_0} & \frac{\partial x_\epsilon}{\partial p_0} \\ \frac{\partial p_\epsilon}{\partial x_0} & \frac{\partial p_\epsilon}{\partial p_0} \end{array} \right] = ?$$

However, L, is a combination of three transformations, $(x_0, p_0) \to (x_0, p_{\epsilon/2})$ then $(x_0, p_{\epsilon/2}) \to (x_\epsilon, p_{\epsilon/2})$ then $(x_\epsilon, p_{\epsilon/2}) \to (x_\epsilon, p_\epsilon)$, each of which has a Jacobian of 1 so $|\det J| = 1$.

Hence the Jacobian of $(x_0, p_0) \to (x_T, p_T)$ is also 1, so

$$f_{X_T, P_T}(x_T(x_0, p_0), p_T(x_0, p_0)) = f_{X_0, P_0}(x_0, p_0) = \pi(x_0) g(p_0; m).$$

where $g(p; m) = (2\pi)^{-1/2} \exp[-p^2/(2m)]$.

$\pi(x)$ is stationary

Since $|\mathrm{det} J| = 1$,

$$\pi(x)q(x'|x) = \pi(x)g(p;m) \propto e^{-H(x,p)}.$$

But, by the skew-symmetry of the leapfrog step (and since $|\mathrm{det} J^{-1}| = 1$)

$$\pi(x')q(x|x') = \pi(x')g(p';m) \propto e^{-H(x',p')}.$$

So, detailed balance is preserved using an acceptance probability of

$$\alpha = \min\left[1, \exp\left\{H(x,p) - H(x',p')\right\}\right].$$

When the leapfrog scheme approximately conserves energy, $\alpha \approx 1$.

Generalising from 1D

Instead of $\partial p/\partial t = -\mathrm{d}U/\mathrm{d}x$ we have

$$\frac{\partial p}{\partial t} = -\nabla U = -\nabla \log \pi.$$

Instead of $\partial x/\partial t = p/m$ we have

$$\frac{\partial x}{\partial t} = M^{-1}p,$$

where M is a (positive definite) mass matrix.

A single leapfrog step then becomes:

$$\begin{aligned} p_{\epsilon/2} &= p_0 - \frac{\epsilon}{2}\nabla \log \pi|_{x_0}, \\ x_\epsilon &= x_0 + \epsilon M^{-1} p_{\epsilon/2}, \\ p_\epsilon &= p_{\epsilon/2} - \frac{\epsilon}{2}\nabla \log \pi|_{x_\epsilon}. \end{aligned}$$

The HMC algorithm

The user chooses a value for tuning parameters T and $\text{nleap} = T/\epsilon$.

Imagine a surface $U(x) = -\log \pi(x)$; our current position x is the position of a ball with mass m.

1. Simulate momentum $P \sim \mathrm{N}(0, m)$).
2. Perform nleap leapfrog steps with a time step of ϵ.
3. New position and momentum, (x', p'), is proposal for next point in Markov chain.
4. Accept proposal with a probability of

   $$\alpha = \min\left(1, \exp\left\{U(x) + \frac{1}{2}p^T M^{-1} p - \left[U(x') + \frac{1}{2}(p')^T M^{-1}(p')\right]\right\}\right);$$

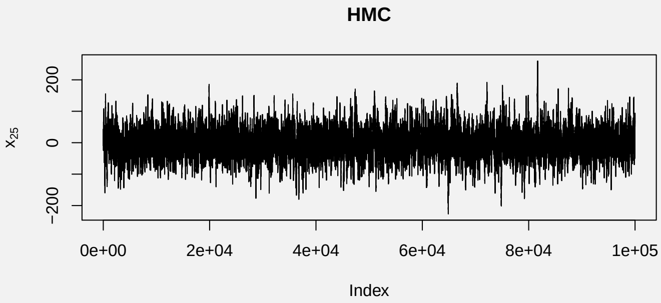
   else stay at x.
5. Go to 1.

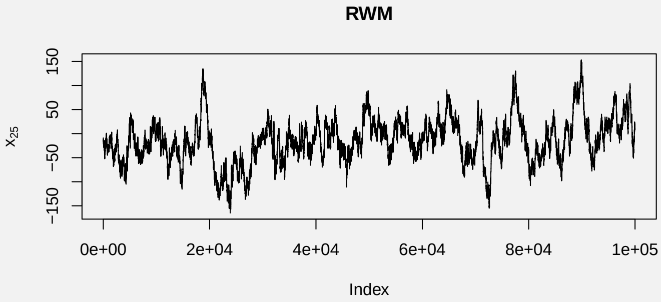
Amazingly, unless ϵ is much too big, the system still approximately conserves energy, even after many leapfrog steps.

Demonstration

Show demonstrations of HMC sampling.

HMC vs RWM





RWM tuned too acceptance rate of $\approx 25\%$; HMC to $\approx 60\%$, using $nleap = 6$. ESSs are ≈ 1840 and ≈ 40.

Problems?

How to choose ϵ? If too small then lots of effort; if too large then energy varies too wildly.

How to choose T? If too small then hardly move; if too large then can almost return to starting position.

Leapfrog includes $p_{\epsilon/2} = p_0 - (\epsilon/2)\nabla \log \pi|_{x_0}$, so the path becomes unstable when the tails are too light; e.g., $\pi(x) \propto e^{-x^4}$.

A single mass matrix may not be appropriate across the whole posterior - but making M position-dependent requires an implicit leapfrog scheme.

Preserve p' - do not throw it away ... but then the algorithm becomes non-reversible!