Missing data lecture 7: Metropolis and Hamiltonian Monte Carlo

MCMC and Gibbs sampling recap

Suppose we want to sample from a distribution $\pi(\theta)$ (for the rest of the lecture I'll suppress the dependence on y unless otherwise noted), but we can't easily do so, we might be able to create a Markov Chain whose stationary distribution is $\pi(\theta)$.

The Markov Chain has the property that

$$P(\theta^{(n)} \in A \mid \theta^{(n-1)} = c_{n-1}, \dots, \theta^{(n-1)} = c_1) = P(\theta^{(n)} \in A \mid \theta^{(n-1)} = c_{n-1})$$

and that

$$P(\theta^{(n)} \in A \mid \theta^{(n-1)} = c)$$

doesn't depend on n.

This transition function has the property that for any value c of $\theta^{(n-1)}$, the function $P(\theta^n \in A \mid \theta^{(n-1)} = c)$ is a probability measure over whatever space A is in, and for any fixed A, $P(\theta^n \in A \mid \theta^{(n-1)} = x)$ is a measurable function of x.

In finite spaces, the transition function is just a matrix with $(i, j)^{\text{th}}$ entry

What the proof from Tanner and Wong (1987) shows is that we can create a Markov Chain with the stationary distribution $p(\theta)$ if we have a transition function $P(\theta \mid \theta')$ with the following properties:

1. $\pi(\theta) = \int_{\theta'} P(\theta \mid \theta')\pi(\theta')d\theta'$
2. $P(\theta \mid \theta') \leq M < \infty$ for all θ, θ'.
3. For every $\theta_0 \in \Omega_\theta$ there is an open neighborhood U of θ_0 so that:

$$P(\theta \mid \theta') > 0, \forall (\theta, \theta') \in U$$

and an initial distribution $g(\theta)$ that satisfies:

$$\sup_\theta g(\theta)/\pi(\theta) < \infty$$

One way of showing condition 1 is by showing that a chain is reversible, namely that for two sets A and B:

$$\int_A \pi(\theta') \int_B p(\theta \mid \theta') d\theta \, d\theta' = \int_B \pi(\theta') \int_A p(\theta \mid \theta') d\theta \, d\theta'$$

We can represent $\int_B p(\theta \mid \theta') d\theta$ as $P(B \mid \theta')$ When A is the whole parameter space, Ω_θ this says something more interpretable:

$$\int_{\Omega_\theta} \pi(\theta') \int_B p(\theta \mid \theta') d\theta \, d\theta' = \int_B \pi(\theta') d\theta'$$

This says: If I draw a value from the stationary distribution, and then draw a value from the transition function, my probability of landing in the set B is the same as if I had just measured whether the first draw was in set B.

Metropolis sampler

This is all from Geyer's notes on MCMC, Geyer (2005).

One way to construct a transition function that has this behavior is by using the Metropolis algorithm.

1. Sample a new point $\theta^{(2)} \mid \theta^{(1)}$ with a proposal distribution that we can draw from, $J(\theta^{(2)} \mid \theta^{(1)})$, such that $J(\theta^{(1)} \mid \theta^{(2)}) = J(\theta^{(2)} \mid \theta^{(1)})$.
2. Accept the new point with probability $\min(r, 1)$:

   $$r = \frac{\pi(\theta^{(2)})}{\pi(\theta^{(1)})}$$

   else set the next step of the Markov Chain to $\theta^{(1)}$.

Crucially, this acceptance probability, which is called the *Metropolis acceptance rate*, can be computed without regards to the normalizing constant of the probability density.

We can see this easily because it's a ratio of the same density evaluated at two different parameters. Let $\pi(\theta) = p(\theta \mid y)$ where $p(\theta)$ is the prior for θ, $f_Y(y \mid \theta)$ is the density of the observations, and $p(y) = \int_\theta p(\theta) f_Y(y \mid \theta) d\theta$:

$$
\begin{aligned}
r &= \frac{p(\theta^{(2)} \mid y)}{p(\theta^{(1)} \mid y)} \\
&= \frac{p(\theta^{(2)}, y)/p(y)}{p(\theta^{(1)}, y)/p(y)} \\
&= \frac{p(\theta^{(2)}, y)}{p(\theta^{(1)}, y)} \\
&= \frac{p(\theta^{(2)}) f_Y(y \mid \theta^{(2)})}{p(\theta^{(1)}) f_Y(y \mid \theta^{(1)})}
\end{aligned}
$$

This is helpful, because we don't know the normalizing constant $p(y)$ for most models we're interested in fitting.

What does the Metropolis algorithm imply for the transition density?

We need to compute the conditional measure $P(A \mid \theta^{(1)})$, which gives the probability of landing in set A, or of drawing a value $\theta^{(2)}$ that is in set A from the algorithm above.

There are two ways we can get to set A. The first way is if the proposed point $\theta^{(2)}$ is in set A and the draw is accepted. The other way is if $\theta^{(1)}$ is in set A and we reject the proposal from $\theta^{(1)} \rightarrow \theta^{(2)}$.

The probability of acceptance for a single point $\theta^{(2)}$ given we started at $\theta^{(1)}$ is $h(\theta^{(1)}, \theta^{(2)}) = \min(r(\theta^{(1)}, \theta^{(2)}), 1)$. The probability that we transition from $\theta^{(1)} \rightarrow \theta^{(2)}$ is given by

$$\int_A J(\theta \mid \theta^{(1)}) h(\theta^{(1)}, \theta) d\theta$$

The probability we accept any jump is the integral over the whole space, Ω_θ:

$$a(\theta^{(1)}) = \int_{\Omega_\theta} J(\theta \mid \theta^{(1)}) h(\theta^{(1)}, \theta) d\theta$$

Then the probability that we reject the proposal at $\theta^{(1)}$ is $1 - a(\theta^{(1)})$. The total probability of landing in set A if we reject the draw is 1 if $\theta^{(1)} \in A$, or 0 if it isn't in A, which we can represent as $\mathbb{1}\left(\theta^{(1)} \in A\right)$.

This means that

$$P(A \mid \theta^{(1)}) = (1 - a(\theta^{(1)}))\mathbb{1}\left(\theta^{(1)} \in A\right) + \int_A h(\theta^{(1)}, \theta) J(\theta \mid \theta^{(1)}) d\theta$$

Now we need to show that this is a reversible transition, namely:

$$\int_B \pi(\theta^{(1)}) P(A \mid \theta^{(1)}) d\theta^{(1)} = \int_A \pi(\theta^{(1)}) P(B \mid \theta^{(1)}) d\theta^{(1)}$$

crucially, for a density $\pi(\theta) \equiv p(\theta \mid y)$ with an unnormalized joint density $p(\theta, y)$ with the property:

$$p(\theta^{(1)}, y)h(\theta^{(1)}, \theta)J(\theta \mid \theta^{(1)}) = p(\theta, y)h(\theta, \theta^{(1)})J(\theta^{(1)} \mid \theta)$$

This is true because, assuming $p(\theta^{(1)}, y) \leq p(\theta, y)$,

$$\begin{aligned} p(\theta^{(1)}, y)h(\theta^{(1)}, \theta)J(\theta \mid \theta^{(1)}) &= p(\theta, y)\frac{p(\theta^{(1)}, y)}{p(\theta, y)}J(\theta^{(1)} \mid \theta) \\ &= p(\theta^{(1)}, y)J(\theta \mid \theta^{(1)}) \end{aligned}$$

$h(\theta^{(1)}, \theta) = \min(p(\theta, y)/p(\theta^{(1)}, y), 1) = 1$. We'll start with the LHS above and we can show that it equals the RHS. First we start with the first term on the LHS:

$$\begin{aligned} \int_B p(\theta^{(1)} \mid y)(1 - a(\theta^{(1)}))\mathbb{1}\left(\theta^{(1)} \in A\right) d\theta^{(1)} &= \int_{\Omega_\theta} \mathbb{1}\left(\theta^{(1)} \in B\right) p(\theta^{(1)} \mid y)(1 - a(\theta^{(1)}))\mathbb{1}\left(\theta^{(1)} \in A\right) d\theta^{(1)} \\ &= \int_{\Omega_\theta} \mathbb{1}\left(\theta^{(1)} \in A\right) p(\theta^{(1)} \mid y)(1 - a(\theta^{(1)}))\mathbb{1}\left(\theta^{(1)} \in B\right) d\theta^{(1)} \\ &= \int_A p(\theta^{(1)} \mid y)(1 - a(\theta^{(1)}))\mathbb{1}\left(\theta^{(1)} \in B\right) d\theta^{(1)}. \end{aligned}$$

$$\begin{aligned} \int_B p(\theta^{(1)} \mid y) \int_A h(\theta, \theta^{(1)})J(\theta \mid \theta^{(1)})d\theta d\theta^{(1)} &= \int_{\Omega_\theta} \mathbb{1}\left(\theta^{(1)} \in B\right) p(\theta^{(1)} \mid y) \int_{\Omega_\theta} \mathbb{1}\left(\theta \in A\right) h(\theta, \theta^{(1)})J(\theta \mid \theta^{(1)})d\theta d\theta^{(1)} \\ &= \int_{\Omega_\theta}\int_{\Omega_\theta} \mathbb{1}\left(\theta \in A\right) \mathbb{1}\left(\theta^{(1)} \in B\right) p(\theta^{(1)} \mid y)h(\theta^{(1)}, \theta)J(\theta \mid \theta^{(1)})d\theta d\theta^{(1)} \\ &= \frac{1}{p(y)}\int_{\Omega_\theta}\int_{\Omega_\theta} \mathbb{1}\left(\theta \in A\right) \mathbb{1}\left(\theta^{(1)} \in B\right) p(\theta^{(1)}, y)h(\theta^{(1)}, \theta)J(\theta \mid \theta^{(1)})d\theta d\theta^{(1)} \\ &= \frac{1}{p(y)}\int_{\Omega_\theta}\int_{\Omega_\theta} \mathbb{1}\left(\theta \in A\right) \mathbb{1}\left(\theta^{(1)} \in B\right) p(\theta, y)h(\theta, \theta^{(1)})J(\theta^{(1)} \mid \theta)d\theta^{(1)} d\theta \\ &= \frac{1}{p(y)}\int_{\Omega_\theta} \mathbb{1}\left(\theta \in A\right) p(\theta, y)\int_{\Omega_\theta} \mathbb{1}\left(\theta^{(1)} \in B\right) h(\theta, \theta^{(1)})J(\theta^{(1)} \mid \theta)d\theta^{(1)} d\theta \\ &= \int_A p(\theta \mid y) \int_B h(\theta, \theta^{(1)})J(\theta^{(1)} \mid \theta)d\theta^{(1)} d\theta \end{aligned}$$

Putting these together, we've shown that:

$$\int_B p(\theta^{(1)} \mid y)P(A \mid \theta^{(1)})d\theta^{(1)} = \int_A \pi(\theta^{(1)} \mid y)P(B \mid \theta^{(1)})d\theta^{(1)}$$

A default Metropolis sampler can be generated using a multivariate normal distribution for $J_t(\theta^b \mid \theta^a)$

$$\theta^b \sim N(\theta^a, \Sigma)$$

where we tune Σ to be about the scale we expect the posterior to be. That means that when we're in regions of high density, we'll have a good chance of jumping to a point that has reasonable posterior density, which means we won't reject the proposal with high probability.

Hamiltonian Monte Carlo

One way to generate a proposal distribution with this property is with an idea from physics using parameter expansion, namely if we have a distribution we'd like to sample from, $\pi(\theta)$, we can introduce 1-to-1 auxiliary variables φ (i.e. if we have d θ, we'll have d φ) with a multivariate normal distribution so our joint target density is $\pi(\theta)\mathcal{N}(\varphi \mid 0, M)$.

If we represent the marginal target density as $\exp(-(-\log \pi(\theta) - \log \mathcal{N}(\varphi \mid 0, M)))$, and call $U(\theta) - \log \pi(\theta)$, $K(\varphi) = -\log \mathcal{N}(\varphi \mid 0, M)))$, we get the following representation:

$$\exp(-(U(\theta) + K(\varphi))) \equiv \pi(\theta)\mathcal{N}(\varphi \mid 0, M)$$

We can think of θ as representing the positions of d particles and φ as representing the momentum. In this sense, $U(\theta)$ is a potential energy, and $K(\varphi)$ is a kinetic energy term. The total energy in the system is $U(\theta) + K(\varphi)$ and this is called the Hamiltonian, or $H(\theta, \varphi)$. It turns out that given an initial starting point (θ_0, φ_0) we can simulate the trajectories of these particles for any time t in the future using the Hamiltonian and what are called Hamilton's system of equations:

$$\frac{d\theta}{dt} = \frac{\partial H(\theta, \varphi)}{\partial \varphi}$$
$$\frac{d\varphi}{dt} = -\frac{\partial H(\theta, \varphi)}{\partial \theta}$$

We can write this in matrix notation if we define the matrix J^{-1} as:

$$\begin{bmatrix} 0 & I_d \\ -I_d & 0 \end{bmatrix}$$

$$\nabla_t \begin{bmatrix} \theta \\ \varphi \end{bmatrix} = J \nabla_{\theta,\varphi} H(\theta, \varphi)$$

Then for a small time step Δt we get

$$\theta_{\Delta t} = \theta_0 + \frac{d\theta}{dt}(\theta, \varphi)\Delta t$$
$$\varphi_{\Delta t} = \varphi_0 + \frac{d\varphi}{dt}(\theta, \varphi)\Delta dt$$

This seems straightforward, but

$$\begin{bmatrix} 0 & -I_d \\ I_d & 0 \end{bmatrix}$$

$$\begin{bmatrix} \theta_{\Delta t} \\ \varphi_{\Delta t} \end{bmatrix} = \begin{bmatrix} \theta \\ \varphi \end{bmatrix} + \begin{bmatrix} 0 & I_d \\ -I_d & 0 \end{bmatrix} \begin{bmatrix} \nabla_\theta H(\theta, \varphi) \\ \nabla_\varphi H(\theta, \varphi) \end{bmatrix} \Delta t$$

What's the Jacobian of this transformation?

$$\nabla_{\theta,\varphi}\begin{bmatrix}\theta_{\Delta t}\\ \varphi_{\Delta t}\end{bmatrix} = \begin{bmatrix}I_d & 0\\ 0 & I_d\end{bmatrix} + \begin{bmatrix}0 & I_d\\ -I_d & 0\end{bmatrix}\begin{bmatrix}\nabla^2_\theta H(\theta,\varphi) & \nabla^2_{\theta,\varphi}H(\theta,\varphi)\\ \nabla^2_{\varphi,\theta}H(\theta,\varphi) & \nabla^2_\varphi H(\theta,\varphi)\end{bmatrix}\Delta t$$

This simplifies to

$$\nabla_{\theta,\varphi}\begin{bmatrix}\theta_{\Delta t}\\ \varphi_{\Delta t}\end{bmatrix} = \begin{bmatrix}I_d + \Delta t\nabla^2_{\theta,\varphi}H(\theta,\varphi) & \Delta t\nabla^2_\theta H(\theta,\varphi)\\ -\Delta t\nabla^2_\varphi H(\theta,\varphi) & I_d - \Delta t\nabla^2_{\theta,\varphi}H(\theta,\varphi)\end{bmatrix}$$

It turns out that the determinant of this matrix is I_d plus terms that involve $(\Delta t)^2$. If we make Δt small, this means that the determinant of the transformation is 1.

Another nice property of these equations is that the Hamiltonian is constant in time:

$$\begin{aligned}\frac{dH(\theta,\phi)}{dt} &= \sum_j \frac{d\theta}{dt}\frac{\partial H}{\partial\theta} + \frac{d\varphi}{dt}\frac{\partial H}{\partial\varphi}\\ &= \sum_j \frac{\partial H}{\partial\varphi}\frac{\partial H}{\partial\theta} - \frac{\partial H}{\partial\theta}\frac{\partial H}{\partial\varphi}\\ &= 0\end{aligned}$$

This means that if we sample (θ_0,φ_0) from the density $\exp(-H(\theta,\varphi))$ and compute the final position and momentum of the particles after t time (θ_t,φ_t), we'll get the same density over θ_0,φ_0 that we started with: Let $F_t(\theta_0,\varphi_0) = (\theta_t,\varphi_t)$. This has an inverse, such that $F_t^{-1}(\theta_t,\varphi_t) = (\theta_0,\varphi_0)$. In fact, this inverse is equal to $F_t^{-1}(\theta_t,\varphi_t) = F_{-t}(\theta_t,\varphi_t)$ Let's compute the density under this transformation, starting with the density:

$$\exp(-H(\theta_0,\varphi_0))d\theta_0 d\varphi_0$$

$$\begin{aligned}\exp(-H(F_t^{-1}(\theta_t\varphi_t)))\det\nabla_{\theta_t,\varphi_t}F_t^{-1}(\theta_t\varphi_t) &= \exp(-H(F_{-t}(\theta_t\varphi_t)))\det\nabla_{\theta_t,\varphi_t}F_t^{-1}(\theta_t\varphi_t)\\ &= \exp(-H(\theta_t\varphi_t))d\theta_t d\varphi_t\end{aligned}$$

where the second line follows from the fact that the change in time for the Hamiltonian is zero, and that the determinant of the transformation is 1. Thus, plugging in θ_t,φ_t to the Hamiltonian has no effect on the value of the function; the Hamiltonian is constant.

This means that the starting values define the total energy for the system.

The algorithm for ex

One key point from above is that we can take gradients without worrying about the normalizing constant!

$$\begin{aligned}\frac{\partial H(\theta,\varphi)}{\partial\theta} &= -\frac{d}{d\theta}\log(p(\theta\mid y))\\ &= -\frac{d}{d\theta}\left(\log p(\theta) + \log(f_Y(y\mid\theta)) - \log(f_Y(y))\right)\\ &= -\frac{d}{d\theta}\left(\log p(\theta) + \log(f_Y(y\mid\theta))\right)\end{aligned}$$

because the marginal density of the data is not dependent on θ. Thus we can use this algorithm to sample from densities with intractable normalizing constants, which is pretty much any interesting statitsical model.

The idea is to draw an initial value for φ from a multivariate normal distribution, and then to run Hamilton's equations to get draws for the final θ^t and φ^t

The problem with this idea is that we can't solve Hamilton's equations for any non-trivial problem. What we do instead is to discretize the equations and solve them approximately. If our Hamiltonian allows for $p(\varphi \mid \theta)$, we'll need to use complex numerical integration schemes. If, as is typical, we use a distribution for φ that is independent of θ (something like a multivariate normal distribution with a fixed covariance matrix, M, for the density $p(\varphi)$), we can use the leapfrog integrator to approximately solve the equations of motion.

Let's define L as the number of steps of the integrator, and ϵ as the step-size, or how finely discretized our equations of motion are. Start with $\theta^{(0)}$, and $\varphi^{(0)} \sim \text{Normal}(0, M)$ and Then for each step $l = 1, \dots, L$, repeat:

$$\begin{aligned}
\varphi^{(\epsilon(l-1/2))} &= \varphi^{(\epsilon(l-1))} + \frac{\epsilon}{2}\frac{d\log(p(\theta \mid y))}{d\theta} \\
\theta^{(\epsilon l)} &= \theta^{(\epsilon(l-1))} + \epsilon M^{-1}\varphi^{(\epsilon(l-1))} \\
\varphi^{(\epsilon l)} &= \varphi^{(\epsilon(l-1/2))} + \frac{\epsilon}{2}\frac{d\log(p(\theta \mid y))}{d\theta}
\end{aligned}$$

Here is the expression for the final proposal for $\theta^{(L\epsilon)}$:

$$\theta^{(L\epsilon)} = \theta^{(0)} + \frac{L\epsilon^2}{2}\nabla_\theta U(\theta^{(0)}) - \epsilon^2\sum_{l=1}^{L}(L-l)\nabla_\theta U(\theta^{(\epsilon l)}) + L\epsilon\varphi^{(0)}$$

where $\varphi^{(0)} \sim \text{Normal}(0, M)$.

One problem with this implementation is that the Hamiltonian isn't exactly conserved, so we do have to do a Metropolis step at the end of the L steps to determine if we accept the proposal.

Finally, after we run the algorithm, we have a new set of parameters $\theta^{(L\epsilon)}, \varphi^{(L\epsilon)}$. We then compute the ratio of the exponentiated Hamiltonian at the start and end of the algorithm:

$$r = \frac{p(\theta^{(0)} \mid y)p(\varphi^{(0)})}{p(\theta^{(L\epsilon)} \mid y)p(\varphi^{(L\epsilon)})}$$

and set $\theta^{(1)} = \theta^{(L\epsilon)}$ with probability $\min(r, 1)$, or $\theta^{(1)} = \theta^{(0)}$ otherwise.

This algorithm has three tuning parameters, L, ϵ, M. One way to set L is to run the algorithm until you detect that the particles have begun to move back towards the starting point, $\theta^{(0)}$. That way, you'll minimize the autocorrelation between draws, and boost your effective sample size.

That suggests a heuristic to measure the dot-product of $(\theta^{(l\epsilon)} - \theta^{(0)})$ and φ. When this becomes negative, it means that the momentum is pointing in a different direction than the difference between the current step and the initial point.

This is the idea behind the No-U-Turn-Sampler, which stops the leapfrog integrator when

$$\sum_j (\theta_j^{(l\epsilon)} - \theta_j^{(0)})\varphi_j^{(l\epsilon)} < 0$$

We can't exactly use this as an exact stopping rule because just choosing $\theta^{(l\epsilon)}$ as the final draw in the trajectory would not lead to a sampler with detailed balance.

See Hoffman, Gelman, et al. (2014) for more information.

Another parameter that needs to be set is ϵ, which is the discretization size of the numerical integrator.

If this is set to be too small, you'll need many leapfrog steps to make measurable progress. If it is set too large, the numerical error in the integrator will add up too quickly and you won't be approximating the solution to the diff-eqs well anymore. This can lead to something called divergences, where the integrator diverges.

Geyer, Charles J. 2005. "Markov Chain Monte Carlo Lecture Notes."

Hoffman, Matthew D, Andrew Gelman, et al. 2014. "The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–623.

Tanner, Martin A, and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398): 528–40.