STA 360/602L: Module 7.1

The Metropolis algorithm

Dr. Olanrewaju Michael Akande

Introduction

- As a refresher, suppose $y = (y_1, \ldots, y_n)$ and each $y_i \sim p(y|\theta)$. Suppose we specify a prior $\pi(\theta)$ on θ.
- Then as usual, we are interested in

$$\pi(\theta|y) = \frac{\pi(\theta)p(y,|\theta)}{p(y)}.$$

- As we already know, it is often difficult to compute $p(y)$.
- Using the Monte Carlo method or Gibbs sampler, we have seen that we don't need to know $p(y)$.
- As long as we have conjugate and semi-conjugate priors, we can generate samples directly from $\pi(\theta|y)$.
- What happens if we cannot sample directly from $\pi(\theta|y)$?

Motivating example

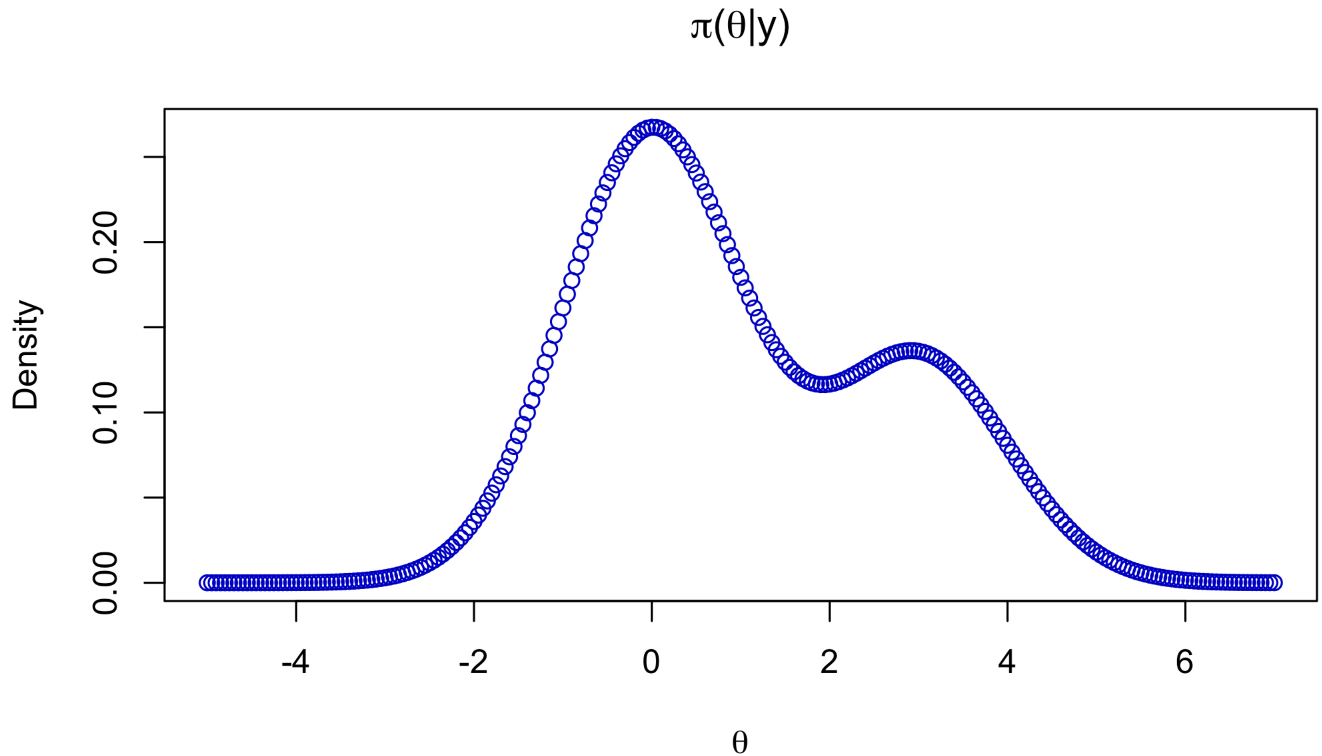
- To motivate our discussions on the Metropolis algorithm, let's explore a simple example.
- Suppose we wish to sample from the following density

$$\pi(\theta|y) \propto \exp^{-\frac{1}{2}\theta^2} + \frac{1}{2}\exp^{-\frac{1}{2}(\theta-3)^2}$$

- This is a *mixture of two normal densities*, one with mode near 0 and the other with mode near 3.
- ***Note****: we will cover finite mixture models properly soon.*
- Anyway, let's use this density to explore the main ideas behind the Metropolis sampler.
- By the way, as you will see, we don't actually need to know the normalizing constant for Metropolis sampling but for this example, find it for practice!

Motivating example

- Let's take a look at the (normalized) density:



- There are other ways of sampling from this density, but let's focus specifically on the Metropolis algorithm here.

Metropolis algorithm

- From a sampling perspective, we need to have a large group of values, $\theta^{(1)}, \ldots, \theta^{(S)}$ from $\pi(\theta|y)$ whose empirical distribution approximates $\pi(\theta|y)$.

- That means that for any two values a and b, we want

$$\frac{\#\theta^{(s)} = a}{S} \div \frac{\#\theta^{(s)} = b}{S} = \frac{\#\theta^{(s)} = a}{S} \times \frac{S}{\#\theta^{(s)} = b} = \frac{\#\theta^{(s)} = a}{\#\theta^{(s)} = b} \approx \frac{\pi(\theta = a|y)}{\pi(\theta = b|y)}$$

- Basically, we want to make sure that if a and b are plausible values in $\pi(\theta|y)$, the ratio of the number of the $\theta^{(1)}, \ldots, \theta^{(S)}$ values equal to them properly approximates $\dfrac{\pi(\theta = a|y)}{\pi(\theta = b|y)}$.

- How might we construct a group like this?

Metropolis algorithm

- Suppose we have a working group $\theta^{(1)}, \ldots, \theta^{(s)}$ at iteration s, and need to add a new value $\theta^{(s+1)}$.

- Consider a candidate value $\theta^\star$ that is close to $\theta^{(s)}$ *(we will get to how to generate the candidate value in a minute)*. Should we set $\theta^{(s+1)} = \theta^\star$ or not?

- Well, we should probably compute $\pi(\theta^\star|y)$ and see if $\pi(\theta^\star|y) > \pi(\theta^{(s)}|y)$. Equivalently, look at $r = \dfrac{\pi(\theta^\star|y)}{\pi(\theta^{(s)}|y)}$.

- By the way, notice that

$$
\begin{aligned}
r = \frac{\pi(\theta^\star|y)}{\pi(\theta^{(s)}|y)} &= \frac{p(y|\theta^\star)\pi(\theta^\star)}{p(y)} \div \frac{p(y|\theta^{(s)})\pi(\theta^{(s)})}{p(y)} \\
&= \frac{p(y|\theta^\star)\pi(\theta^\star)}{p(y)} \times \frac{p(y)}{p(y|\theta^{(s)})\pi(\theta^{(s)})} = \frac{p(y|\theta^\star)\pi(\theta^\star)}{p(y|\theta^{(s)})\pi(\theta^{(s)})},
\end{aligned}
$$

which does not depend on the marginal likelihood we don't know!

Metropolis algorithm

- If $r > 1$
  - Intuition: $\theta^{(s)}$ is already a part of the density we desire and the density at $\theta^\star$ is even higher than the density at $\theta^{(s)}$.
  - Action: set $\theta^{(s+1)} = \theta^\star$
- If $r < 1$,
  - Intuition: relative frequency of values on our group $\theta^{(1)}, \ldots, \theta^{(s)}$ equal to $\theta^\star$ should be $\approx r = \dfrac{\pi(\theta^\star|y)}{\pi(\theta^{(s)}|y)}$. For every $\theta^{(s)}$, include only a fraction of an instance of $\theta^\star$.
  - Action: set $\theta^{(s+1)} = \theta^\star$ with probability r and $\theta^{(s+1)} = \theta^{(s)}$ with probability $1 - r$.

Metropolis algorithm

- This is the basic intuition behind the Metropolis algorithm.
- Where should the proposed value $\theta^\star$ come from?
- Sample $\theta^\star$ close to the current value $\theta^{(s)}$ using a symmetric proposal distribution $g[\theta^\star|\theta^{(s)}]$. g is actually a "family of proposal distributions", indexed by the specific value of $\theta^{(s)}$.
- Here, symmetric means that $g[\theta^\star|\theta^{(s)}] = g[\theta^{(s)}|\theta^\star]$.
- The symmetric proposal is usually very simple with density concentrated near $\theta^{(s)}$, for example, $\mathcal{N}(\theta^\star; \theta^{(s)}, \delta^2)$ or $\mathrm{Unif}(\theta^\star; \theta^{(s)} - \delta, \theta^{(s)} + \delta)$.
- After obtaining $\theta^\star$, either add it or add a copy of $\theta^{(s)}$ to our current set of values, depending on the value of r.

Metropolis algorithm

- The algorithm proceeds as follows:

1. Given $\theta^{(1)}, \ldots, \theta^{(s)}$, generate a *candidate* value $\theta^\star \sim g[\theta^\star | \theta^{(s)}]$.

2. Compute the acceptance ratio

$$r = \frac{\pi(\theta^\star | y)}{\pi(\theta^{(s)} | y)} = \frac{p(y|\theta^\star)\pi(\theta^\star)}{p(y|\theta^{(s)})\pi(\theta^{(s)})}.$$

3. Set

$$\theta^{(s+1)} = \begin{cases} \theta^\star & \text{with probability} \quad \min(r, 1) \\ \theta^{(s)} & \text{with probability} \quad 1 - \min(r, 1) \end{cases}$$

which can be accomplished by sampling $u \sim U(0,1)$ independently and setting

$$\theta^{(s+1)} = \begin{cases} \theta^\star & \text{if} \quad u < r \\ \theta^{(s)} & \text{if} \quad \text{otherwise} \end{cases}.$$

Metropolis algorithm

- Once we obtain the samples, then we are back to using Monte Carlo approximations for quantities of interest.
- That is, we can again approximate posterior means, quantiles, and other quantities of interest using the empirical distribution of our sampled values.
- *Some notes:*
  - The Metropolis chain ALWAYS moves to the proposed $\theta^\star$ at iteration $s + 1$ if $\theta^\star$ has higher target density than the current $\theta^{(s)}$.
  - Sometimes, it also moves to a $\theta^\star$ value with lower density in proportion to the density value itself.
  - This leads to a random, Markov process than naturally explores the space according to the probability defined by $\pi(\theta|y)$, and hence generates a sequence that, while dependent, eventually represents draws from $\pi(\theta|y)$.

Metropolis algorithm: convergence

- We will not cover the convergence theory behind Metropolis chains in detail, but below are a few notes for those interested:
  - The Markov process generated under this condition is ergodic and has a limiting distribution.
  - Here, think of ergodicity as meaning that the chain can move anywhere at each step, which is ensured, for example, if $g[\theta^\star|\theta^{(s)}] > 0$ everywhere!
  - By construction, it turns out that the Metropolis chains are reversible, so that convergence to $\pi(\theta|y)$ is assured.
  - Think of reversibility as being equivalent to symmetry of the joint density of two consecutive $\theta^{(s)}$ and $\theta^{(s+1)}$ in the stationary process, which we do have by using a symmetric proposal distribution.
- If you want to learn more about convergence of MCMC chains, consider taking one of the courses on stochastic processes, or Markov chain theory.

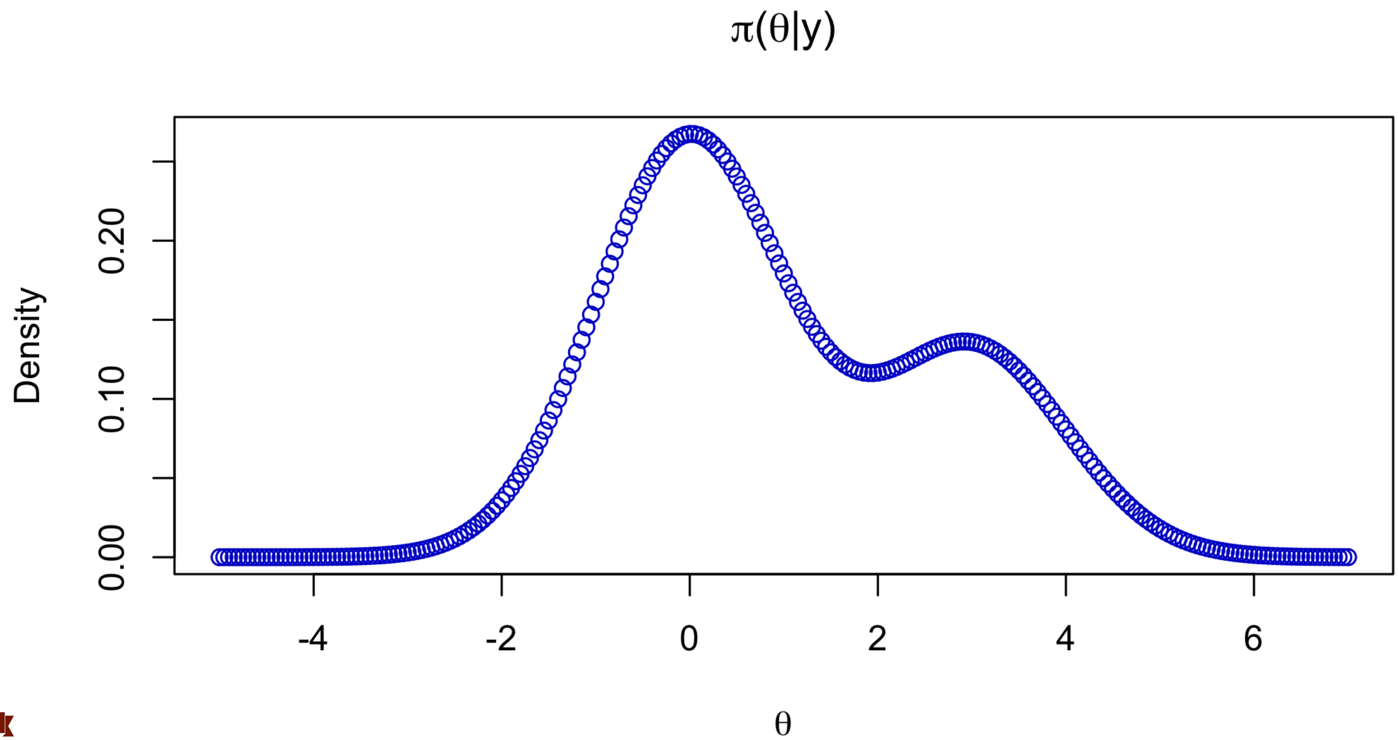
Metropolis algorithm: tuning

- Correlation between samples can be adjusted by selecting optimal δ (i.e., spread of the distribution) in the proposal distribution
- Decreasing correlation increases the effective sample size, increasing rate of convergence, and improving the Monte Carlo approximation to the posterior.
- However,
  - δ too small leads to $r \approx 1$ for most proposed values, a high acceptance rate, but very small moves, leading to highly correlated chain.
  - δ too large can get "stuck" at the posterior mode(s) because $\theta^{\star}$ can get very far away from the mode, leading to a very low acceptance rate and again high correlation in the Markov chain.
- Thus, good to implement several short runs of the algorithm varying δ and settle on one that yields acceptance rate in the range of 25-50%.
- Burn-in (and thinning) is even more important here!

Metropolis in action

Back to our example with

$$\pi(\theta|y) \propto \exp^{-\frac{1}{2}\theta^2} + \frac{1}{2}\exp^{-\frac{1}{2}(\theta-3)^2}$$



Move to the R script here.

What's next?

Move on to the readings for the next module!