1 Outline proof of convergence of Metropolis sampler to the target posterior distribution (based on Gelman et al: BDA, 3rd edn pp 279-80).

1.1 Preliminaries

Suppose we are interested in the posterior distribution

$$\begin{aligned} p(\boldsymbol{\theta}|\text{data}) &= \frac{p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\theta} \\ &\propto p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\text{data}) \end{aligned}$$

and propose to approximate the distribution using a Metropolis (or Metropolis - Hastings algorithm)

Let $J_t(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{(t-1)})$ denote a symmetric jumping or proposal distribution from which a jump from the value of the $\boldsymbol{\theta}$ at the $(t-1)^{th}$ iteration of the chain is generated. In this context symmetry means $J_t(\boldsymbol{\theta}^{new}|\boldsymbol{\theta}^{(t-1)}) = J_t(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{new})$ for all values of $\boldsymbol{\theta}^{new}$ and $\boldsymbol{\theta}^{(t-1)}$. For the Metropolis-Hastings algorithm the requirement that $J_t()$ be symmetric is relaxed.

1. Assume the sequence of parameter values $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)})$ constitute a Markov chain, i.e.

   $$p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(0)}) = p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}) \tag{1}$$

   This holds by construction. (Here superscripts are iteration indices, not powers)

2. Assume the chain is

   (a) aperiodic (no deterministic movement)

   (b) not transient (positive recurrent– some chance of returning to the initial value);

(c) irreducible (the jumping distribution $J_t()$ is able eventually to jump to any state).

3. The above conditions imply the Markov Chain has a unique stationary distribution.

4. Hence the the marginal distributions from which the $\boldsymbol{\theta}$ values are drawn converge to a unique stationary distribution.

5. If we can show that the stationary distribution of the Markov Chain is the target posterior distribution $p(\boldsymbol{\theta}|\text{data})$ then we are done.

Definition of stationarity
if $g(\boldsymbol{\theta})$ is some density such that if $\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{(t)}$ are successive values from the Markov Chain $\tilde{\boldsymbol{\theta}}$,

$$p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a) = g(\boldsymbol{\theta}_a) \Rightarrow p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a) = g(\boldsymbol{\theta}_a) \text{ for all } \boldsymbol{\theta}_a \tag{2}$$

then g is a stationary distribution of the Markov Chain. That is if $\boldsymbol{\theta}^{(t-1)}$ being a draw from the density $g()$ *implies* $\boldsymbol{\theta}^{(t)}$ is also a draw from the density $g()$ then $g()$ is a stationary distribution of the Markov Chain. Here, I am using notation such as $p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a)$ to refer to a density function for $\boldsymbol{\theta}^{(t)}$ evaluated at the specific point $\boldsymbol{\theta}_a$ (so $p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a)\,\forall \boldsymbol{\theta}_a$ refers to the full density function.)

1.2 Outline Proof that the posterior distribution is the stationary distribution of the chain

Suppose the chain is aperiodic, not transient and irreducible and has been running for some time. Consider the chain at the $(t-1)^{th}$ iteration and suppose that the marginal distribution of $\boldsymbol{\theta}^{(t-1)}$ is the target distribution $p(\boldsymbol{\theta}|\text{data})$, that is $p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a) = p(\boldsymbol{\theta} = \boldsymbol{\theta}_a|\text{data})$, for all $\boldsymbol{\theta}_a$. If we can show that this implies the marginal distribution of $p(\boldsymbol{\theta}^{(t)})$ is also $p(\boldsymbol{\theta}|\text{data})$ meaning $p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a) = p(\boldsymbol{\theta} = \boldsymbol{\theta}_a|\text{data})\,\forall \boldsymbol{\theta}_a$ then we are done. In fact we need only show that the $(t-1)^{th}$ draw in the chain is a draw from the target posterior distribution implies $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)})$ are exchangeable, since the marginals of exchangeable r.v's are identical (proved in Assignment 4). Recall that exchangeability means

$$p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a, \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_b) = p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_b, \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a), \text{ for all } \boldsymbol{\theta}_a, \boldsymbol{\theta}_b$$

We will sometimes just write $p(\boldsymbol{\theta}_a|\text{data})$ to refer to the value of the posterior density function for $\boldsymbol{\theta}$ evaluated at the point $\boldsymbol{\theta}_a$, i.e $p(\boldsymbol{\theta}_a|\text{data})$ is shorthand for $p(\boldsymbol{\theta} = \boldsymbol{\theta}_a|\text{data})$.

Let $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ be arbitrary points drawn from the target posterior density but labelled so that

$$p(\boldsymbol{\theta}_b|\text{data}) \geq p(\boldsymbol{\theta}_a|\text{data}) \tag{3}$$

implying that the Metropolis acceptance ratio for a jump from $\boldsymbol{\theta}_a$ to $\boldsymbol{\theta}_b$ is

$$r_M(\boldsymbol{\theta}_b, \boldsymbol{\theta}_a) = \frac{p(\boldsymbol{\theta}_b|\text{data})}{p(\boldsymbol{\theta}_a|\text{data})} = \frac{q(\boldsymbol{\theta}_b|\text{data})}{q(\boldsymbol{\theta}_a|\text{data})} \geq 1. \tag{4}$$

Consequently, a jump from $\boldsymbol{\theta}_a$ to $\boldsymbol{\theta}_b$ will always be accepted, since the acceptance probability for a proposed jump in that direction is $\min\left(1, r_M(\boldsymbol{\theta}_b, \boldsymbol{\theta}_a)\right)$.

Note in (4), the second equality follows because the normalising constant cancels from the numerator and denominator of the ratio of posterior densities. As usual, we need only be able to compute the unnormalized posterior densities.

On the other hand, the Metropolis acceptance ratio for a proposed jump from θ_b to θ_a is

$$r_M(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) = \frac{p(\boldsymbol{\theta}_a|\text{data})}{p(\boldsymbol{\theta}_b|\text{data})} = \frac{q(\boldsymbol{\theta}_a|\text{data})}{q(\boldsymbol{\theta}_b|\text{data})} < 1, \tag{5}$$

so a jump from $\boldsymbol{\theta}_b$ to $\boldsymbol{\theta}_a$ will be accepted with probability $r_M(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$. The joint probability density for observing $\boldsymbol{\theta}_a$ at the $(t-1)^{th}$ iteration and $\boldsymbol{\theta}_b$ at the t^{th} iteration is

$$\begin{aligned}
p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a, \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_b) &= p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a) p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_b | \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a) \\
&= p(\boldsymbol{\theta} = \boldsymbol{\theta}_a|\text{data}) J_t(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a) \times \min\left(1, r_M(\boldsymbol{\theta}_b, \boldsymbol{\theta}_a)\right) && (6) \\
&= p(\boldsymbol{\theta} = \boldsymbol{\theta}_a|\text{data}) J_t(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a) && (7) \\
&= p(\boldsymbol{\theta}_a|\text{data}) J_t(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a) && (8)
\end{aligned}$$

since $r_M(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) \geq 1$ because of the way we have arranged $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$. In (6) I have used the assumption that $\boldsymbol{\theta}^{(t-1)}$ is a draw from $p(\boldsymbol{\theta}|\text{data})$.

The joint probability density for observing $\boldsymbol{\theta}_b$ at iteration $(t-1)$ and $\boldsymbol{\theta}_a$ at iteration (t) is

$$\begin{aligned}
p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_b, \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a) &= p(\boldsymbol{\theta}_b|\text{data}) p(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a | \boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_b) \\
&= p(\boldsymbol{\theta}_b|\text{data}) J_t(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) \times \min\left(1, r_M(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)\right) \\
&= p(\boldsymbol{\theta}_b|\text{data}) J_t(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) \frac{p(\boldsymbol{\theta}_a|\text{data})}{p(\boldsymbol{\theta}_b|\text{data})} \\
&= p(\boldsymbol{\theta}_a|\text{data}) J_t(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) \\
&= p(\boldsymbol{\theta}_a|\text{data}) J_t(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a) \qquad (9)
\end{aligned}$$

where (9) follows because of the symmetry of $J_t()$. Thus, comparing (8) and (9) it is clear that

$$p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_a, \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_b) = p(\boldsymbol{\theta}^{(t-1)} = \boldsymbol{\theta}_b, \boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}_a)$$

and this holds for any $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$. That is $(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{(t)}$ are exchangeable hence

$$p(\boldsymbol{\theta}^{(t)}) = p(\boldsymbol{\theta}^{(t-1)}) = p(\boldsymbol{\theta}|\text{data})$$

since we assumed $\boldsymbol{\theta}^{(t-1)}$ is a draw from the posterior distribution.

Thus we have shown $p(\boldsymbol{\theta}^{(t-1)})$=$p(\boldsymbol{\theta}|\text{data})$ implies $p(\boldsymbol{\theta}^{(t)} = p(\boldsymbol{\theta}|\text{data})$ In other words, the posterior distribution is a stationary distribution of the Markov Chain; In view of the uniqueness of the stationary distribution the chain converges to the posterior distribution.

In the case of the Metropolis-Hastings algorithm the proof is very similar because the acceptance probabilities adjust for the assymetry in $J_t()$.