# Basic MCMC

Andrew Brown[1]

SAMSI

[1] Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA

# Gibbs Sampling

- Suppose we have a $k$-dimensional parameter space, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, from which we wish to sample the posterior distribution, $\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.
- Assuming the availability of the *full conditional* distributions, $\{p(\theta_i \mid \boldsymbol{\theta}_{(-i)}, \mathbf{y})\ i = 1, \ldots, k\}$, the Gibbs sampling algorithm is as follows:

  1 Draw $\theta_1^{(t)} \sim p(\theta_1 \mid \theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)})$
  2 Draw $\theta_2^{(t)} \sim p(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_k^{(t-1)})$
  $\vdots$
  $k$ Draw $\theta_k^{(t)} \sim p(\theta_k \mid \theta_1^{(t)}, \ldots, \theta_{k-1}^{(t)})$

  Repeat for $t = 1, \ldots, T$.
- Under "mild" conditions, $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_k^{(t)})^T$ converges to a draw from $p(\theta_1, \ldots, \theta_k \mid \mathbf{y})$

Geman and Geman (1984), Carlin and Louis (2009)

- After a sufficiently large number of iterations (the “burn in” period), the simulated draws may be treated as realizations from the posterior of interest
- Common practice is to run multiple chains (3 or 5) starting from different values and compare the draws from each chain (after burn-in) to assess convergence
  - Chains can be run in parallel on a computer to save processor time
  - Can also use quantitative measures such as the *scale reduction factor*
  - How to assess convergence with high-dimensional parameter spaces is not clear (look at draws of a few selected parameters, realizations of the likelihood function, etc.)
- The parameters can be partitioned into subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \ldots, \boldsymbol{\theta}_P^T)^T$ and updated in blocks $p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_P, \mathbf{y})$, etc.
  - This can speed up convergence when subsets of the parameters are highly correlated in the posterior

Gelman and Rubin (1992)

# Metropolis Algorithm

- Target distribution with density $h(\boldsymbol{\theta})$
- Proposal density: $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)$
- Metropolis Algorithm:
  1. Draw $\boldsymbol{\theta}^* \sim q(\cdot \mid \boldsymbol{\theta}^{(t-1)})$.
  2. Set $r = h(\boldsymbol{\theta}^*)/h(\boldsymbol{\theta}^{(t-1)}) = \exp(\log h(\boldsymbol{\theta}^*) - \log h(\boldsymbol{\theta}^{(t-1)}))$
  3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$; else set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ w.p. $r$, $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ otherwise.

  Repeat for $t = 1, \ldots, T$
- Metropolis-Hastings Algorithm: Use a candidate density so that $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) \neq q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)$ and set the acceptance ratio to

$$r = \frac{h(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^{(t-1)}) q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}$$

- Gibbs sampling can be shown to be a special case of MH with acceptance probability = 1.

Metropolis et al. (1953), Hastings (1970), Chib and Greenberg (1995), Gelman et al. (2014)

# Example (Chib and Greenberg, 1995)

- Target distribution: $N_2\left(\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right)$
- Metropolis proposals:
  1. Proposal distribution = Uniform $\left([\theta_1^{(t-1)} - 1, \theta_1^{(t-1)} + 1] \times [\theta_2^{(t-1)} - 1.5, \theta_2^{(t-1)} + 1.5]\right)$
  2. Proposal distribution = $N_2\left(\boldsymbol{\theta}^{(t-1)}, \begin{pmatrix}0.6 & 0\\0 & 0.4\end{pmatrix}\right)$
- Gibbs sampling
  - $\theta_1^{(t)} \mid \theta_2^{(t-1)} \sim N(\mu_1 + 0.9(\theta_2^{(t-1)} - \mu_2),\ 1 - 0.9^2)$
  - $\theta_2^{(t)} \mid \theta_1^{(t)} \sim N(\mu_2 + 0.9(\theta_1^{(t)} - \mu_1),\ 1 - 0.9^2)$

# Random Walk Metropolis

- Most of the work in Metropolis and MH algorithms is in finding a good proposal density, $q$, so that the chain efficiently explores the posterior support
  - Proposed jumps too big $\Rightarrow$ few candidates accepted $\Rightarrow$ takes a long time to converge
  - Proposed jumps too small $\Rightarrow$ chain barely moves around the parameter space $\Rightarrow$ takes a long time to converge
- Common to transform $\boldsymbol{\theta}$ to remove boundary constraints (if necessary) and use a Gaussian proposal centered around $\boldsymbol{\theta}^{(t-1)}$:

$$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}, \Sigma)$$

- Choose $\Sigma$ so that:
  - (Univariate case, $\dim(\boldsymbol{\theta}) = 1$): Acceptance rate $\approx 44.1\%$
  - (Multivariate case, $\dim(\boldsymbol{\theta}) \to \infty$): Acceptance rate $\approx 23.4\%$.
- $\Sigma$ can be tuned *adaptively* with $\Sigma = c \cdot \Sigma_0$, where $c$ is increased or decreased throughout the burn-in period based on the monitored acceptance rates

Carlin and Louis (2009)

- Implementing Gibbs sampling directly requires the ability to draw from each full conditional distribution
- In many practical Bayesian models (including the ones discussed here), not all full conditional distributions are available
- We can use other sampling techniques on the individual conditional distributions to approximate the necessary draws
    - “Metropolis-within-Gibbs”: Using a Metropolis updating step on one or more full conditional distributions inside a Gibbs sampling algorithm

# Slice Sampling

- Target distribution: $p(\theta) \propto f(\theta)$, $f(\theta)$ known
- Introduce auxiliary variable $U$ with $U \mid \theta \sim \text{Unif}(0, f(\theta))$. Then

$$p(\theta, u) \propto f(\theta) \frac{1}{f(\theta)} \cdot I(U < f(\theta)) = I(U < f(\theta)).$$

- Gibbs sampling from this joint distribution requires
  - (i) $U \mid \theta \sim \text{Unif}(0, f(\theta))$
  - (ii) $\theta \mid U \sim \text{Unif}(\{\theta : f(\theta) \geq U\})$
- (i) is easy; (ii) can be difficult

Neal (2003), Carlin and Louis (2009)

Illustration: To move from $x_0$ to $x_1$,

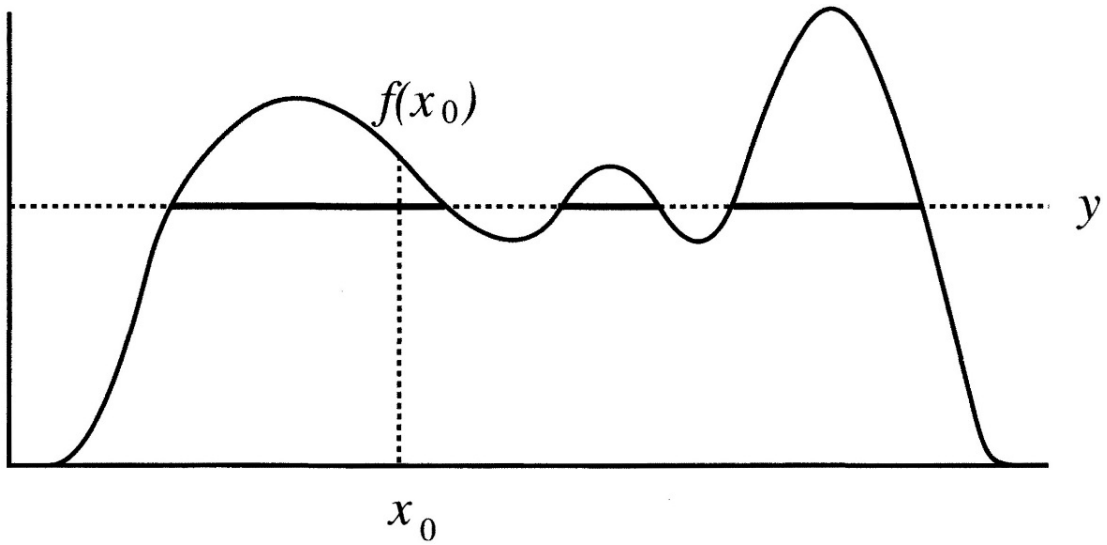
(a) Draw $y$ uniformly from $(0, f(x_0))$, "slicing" the density with $S = \{x : y < f(x)\}$.

(b) Find an interval $I = (L, R)$ around $x_0$ that contains all (or most) of $S$

(c) Draw $x_1$ uniformly from $S \cap I$

(d) $x_1$ is acceptable if it is in the set

$$A = \{x : x \in S \cap I \text{ and } P(\text{Choose } I \mid \text{Starting from } x) = P(\text{Choose } I \mid \text{Starting from } x_0)\}$$




Finding an appropriate $I$ can be tricky, since we don't want $|I|$ to be much bigger than $|S|$.

Neal (2003)

# Slice Sampling vs. Metropolis-Hastings

- Slice sampling draws from the *exact* distribution $p(\theta)$, whereas M-H only *approximates* a draw from this distribution
  - Slice sampler inside a Gibbs algorithm converges more rapidly (in terms of necessary number of draws)
- It can take many candidate draws of $\theta$ before one is retained, so M-H may be better in terms of overall computation time.

# Langevin-Hastings Algorithm

- Target distribution with density $h(\boldsymbol{\theta})$:
- For $t = 1, \ldots, T$, repeat:
  1. Draw $\boldsymbol{\theta}^* \sim N(\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}^{(t-1)}),\ \tilde{\Sigma})$, where

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}^{(t-1)}) = \boldsymbol{\theta}^{(t-1)} + \frac{\sigma^2}{2} \nabla \log h(\boldsymbol{\theta}^{(t-1)}),$$

     and

$$\tilde{\Sigma} = \sigma^2 \mathbf{I}.$$

  2. Acceptance ratio:

$$r = \frac{h(\boldsymbol{\theta}^*) \exp\left[-\frac{1}{2\sigma^2} \|\boldsymbol{\theta}^{(t-1)} - \boldsymbol{\theta}^* - \frac{\sigma^2}{2} \nabla \log h(\boldsymbol{\theta}^*)\|^2\right]}{h(\boldsymbol{\theta}^{(t-1)}) \exp\left[-\frac{1}{2\sigma^2} \|\boldsymbol{\theta}^* - \boldsymbol{\theta}^{(t-1)} - \frac{\sigma^2}{2} \nabla \log h(\boldsymbol{\theta}^{(t-1)})\|^2\right]}$$

  3. If $r \geq 1$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$; else set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$ w.p. $r$, $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$ otherwise.
- Tune $\sigma^2$ to obtain acceptance rate $\approx 57.4\%$.

Grenader and Miller (1994), Gilks et al. (1996), Roberts and Rosenthal (1998, 2001) Carlin and Louis (2009)