

# Advanced Bayesian Computing: Prereqs Review

Andrew J. Holbrook

UCLA Biostatistics 285

Spring 2021

# Bayesian inference

We observe data  $y_1, \dots, y_N \stackrel{iid}{\sim} p(y_n|\boldsymbol{\theta})$  and assume  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ .  
Here,

- ▶  $p(y|\boldsymbol{\theta}) = \prod_{n=1}^N p(y_n|\boldsymbol{\theta})$  is the *likelihood*,
- ▶  $p(\boldsymbol{\theta})$  is the *prior*,

and the goal of Bayesian inference is to obtain the *posterior*

$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y)} = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

# Bayesian inference

We're usually interested in computing another integral

$$\mathbb{E}_{\theta|y} f(\theta) = \int_{\Theta} f(\theta) p(\theta|y) d\theta,$$

so we do what statisticians have been doing forever. We collect samples and rely on the law of large numbers. Suppose  $\theta_1, \dots, \theta_S \stackrel{iid}{\sim} p(\theta|y)$  ( $\mathbb{E}_{\theta|y} |\theta| < \infty$ ) and  $f(\cdot)$  a.s. continuous, then

- ▶ (WLLN)  $\sum_{s=1}^S f(\theta_s)/S \xrightarrow{P} \mathbb{E}_{\theta|y} f(\theta)$
- ▶ (SLLN)  $\sum_{s=1}^S f(\theta_s)/S \xrightarrow{a.s.} \mathbb{E}_{\theta|y} f(\theta)$

But where do we find our samples?

# Generating (pseudo) random variables

We want to sample  $Y \sim F(y)$ , where  $F(\cdot)$  is the (monotonically increasing) c.d.f.

Claim 1

*Assume we can generate  $U \sim U(0, 1)$  and compute  $F^{-1}(\cdot)$ . Then*

$$F^{-1}(U) \sim F(y).$$

Proof.

$$\Pr(F^{-1}(U) < y) = \Pr(U < F(y)) = F(y).$$



# Exponential random variables

Ingredients for  $Y \sim \exp(\lambda)$ :

1.  $p(Y|\lambda) = \lambda \exp(-\lambda Y)$

2.  $F(y|\lambda) = \Pr(Y < y|\lambda) = \int_0^y \lambda \exp(-\lambda Y) = 1 - \exp(-\lambda Y)$

3.  $F^{-1}(u) = -\lambda^{-1} \log(1 - u)$

Easy but *extremely* limited!

## Part 1. Monte Carlo

# Rejection sampling *Von Neumann, 1951*

We want to sample from generic  $p(\theta)$  but only know  $c \cdot p^*(\theta) = p(\theta)$   
 $p^*(\theta) \propto p(\theta)$ . We can easily sample from  $q(\theta)$  and know a  
number  $M > 0$  s.t.  $p^*(\theta) < \underline{Mq(\theta)}$ .

Algorithm for generating  $\theta \sim p(\theta)$ :

1. Draw  $\underline{\theta^*} \sim q(\theta)$  and  $U \sim U(0, 1)$
2.  $\theta \leftarrow \theta^*$  if  $U < \frac{p^*(\theta)}{Mq(\theta)}$

The tighter the envelope  $Mq(\theta)$ , the better. Suppose  
 $q(\theta) = p(\theta) = c \cdot p^*(\theta)$ . Then

$$\Pr(\text{Accept}) = \frac{1}{c \cdot M},$$

and expected number of iterations for one sample is  $c \cdot M$ .

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{p^*(\theta)}{Mq(\theta)} q(\theta) d\theta \\ &= \int_{-\infty}^{\infty} \frac{p^*(\theta)}{c \cdot M} d\theta \\ &= \frac{1}{c \cdot M} \end{aligned}$$

Validity of rejection sampling

$$K(\theta) = \frac{p^*(\theta)}{Mg(\theta)}, \quad p(\theta) = p^*(\theta) c^0$$

$$\text{WTS } \Pr(\theta < v | \text{Accept}) = \int_{-\infty}^v p(\theta) d\theta$$

---

$$\text{LHS} = \Pr(\theta < v | \text{Accept}) / \Pr(\text{Accept})$$

$$= \int_{-\infty}^v \cancel{K(\theta)} \cancel{g(\theta)} d\theta / \int_{-\infty}^{\infty} K(\theta) g(\theta) d\theta \quad \frac{M}{M}$$

$$= \int_{-\infty}^v \cancel{p^*(\theta)} d\theta / \int_{-\infty}^{\infty} \cancel{p^*(\theta)} d\theta \quad \frac{c^0}{c^0}$$

$= \text{RHS}$



# Importance sampling

We wish to know  $\mathbb{E}f(\boldsymbol{\theta}) = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ . We can evaluate  $p^*(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})$  and can sample from  $q(\boldsymbol{\theta})$  easily.

Algorithm for generating estimator  $\widehat{\mathbb{E}f(\boldsymbol{\theta})}$ :

1. Draw  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_S \sim q(\boldsymbol{\theta})$
2. Calculate  $w_k = \frac{w(\boldsymbol{\theta}_k)}{\sum_{s=1}^S w(\boldsymbol{\theta}_s)}$ ,  $w(\boldsymbol{\theta}_s) = \frac{p^*(\boldsymbol{\theta}_s)}{q(\boldsymbol{\theta}_s)}$  for  $k = 1, \dots, S$ .
3. Return  $\sum_{s=1}^S w_s f(\boldsymbol{\theta}_s)$

# Validity of importance sampling

By the LLN,

$$\frac{1}{S} \sum_{s=1}^S w(\theta_s) f(\theta_s) \xrightarrow{a.s.} \int w(\theta) f(\theta) q(\theta) d\theta \quad \text{and}$$

$$\frac{1}{S} \sum_{s=1}^S w(\theta_s) \xrightarrow{a.s.} \int w(\theta) q(\theta) d\theta.$$

Therefore,

$$\begin{aligned} \sum_{s=1}^S w_s f(\theta_s) &= \frac{\frac{1}{S} \sum_{s=1}^S w(\theta_s) f(\theta_s)}{\left[ \frac{1}{S} \sum_{s=1}^S w(\theta_s) \right]} \xrightarrow{a.s.} \frac{\int w(\theta) f(\theta) q(\theta) d\theta}{\int w(\theta) q(\theta) d\theta} \\ &= \frac{\int f(\theta) p^*(\theta) d\theta}{\int p^*(\theta) d\theta} = \int f(\theta) p(\theta) d\theta = \mathbb{E}f(\theta). \end{aligned}$$

substituting

$$w(\theta) = \frac{p(\theta)}{q(\theta)}$$

↓  
I

∩

## Variance of IS estimator

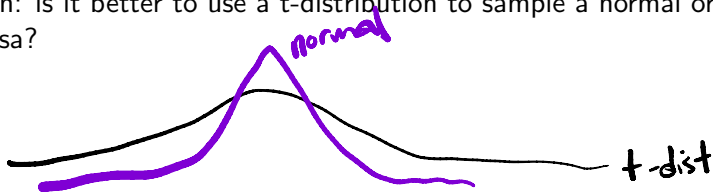
$$w(\theta) = p^*(\theta) / q(\theta)$$

An estimator for the variance of  $\widehat{\mathbb{E}f(\theta)} = \sum_{s=1}^S w_s f(\theta_s)$  is

$$\widehat{\text{Var}}\left(\widehat{\mathbb{E}f(\theta)}\right) \approx \sum_{s=1}^S w_s^2 (f(\theta_s) - \widehat{\mathbb{E}f(\theta)})^2.$$

The variance can be large if even a single  $w_s$  is large.

Question: is it better to use a t-distribution to sample a normal or vice-versa?



Part 2. Discrete time, discrete space,  
time-homogeneous Markov chains

# The setup

Our Markov chain is a discrete time stochastic process  $\{\boldsymbol{\theta}^{(s)}, s \in \mathbb{N}\}$  satisfying

$$\Pr(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^{(s-2)}, \dots, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(0)}) = \Pr(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}).$$

Ingredients:

1. The state space  $\mathcal{S}$  is a finite or countable set.
2. Initial distribution  $\{p_i^{(0)}\}_{i \in \mathcal{S}}$ , satisfying
  - 2.1  $p_i^{(0)} = \Pr(\boldsymbol{\theta}^{(0)} = i)$
  - 2.2  $p_i^{(0)} \geq 0$
  - 2.3  $\sum_{i \in \mathcal{S}} p_i^{(0)} = 1$
3. Transition probabilities  $\{q_{ij}\}_{i, j \in \mathcal{S}}$ 
  - 3.1  $q_{ij} = \Pr(\boldsymbol{\theta}^{(s)} = j | \boldsymbol{\theta}^{(s-1)} = i)$
  - 3.2  $q_{ij} \geq 0$
  - 3.3  $\sum_{j \in \mathcal{S}} q_{ij} = 1$

## Finite state space

When  $\mathcal{S} = \{1, \dots, M\}$ , then we can write state probabilities as row-vectors:

$$\mathbf{p}^{(s)} = \left( \Pr(\boldsymbol{\theta}^{(s)} = 1), \Pr(\boldsymbol{\theta}^{(s)} = 2), \dots, \Pr(\boldsymbol{\theta}^{(s)} = M) \right)$$

Similarly, the transition probabilities  $q_{ij}$  form the matrix

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & \dots & q_{1M} \\ q_{21} & q_{22} & \dots & q_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ q_{M1} & q_{M2} & \dots & q_{MM} \end{bmatrix}$$

and

$$\mathbf{p}^{(s)} = \mathbf{p}^{(s-1)}\mathbf{Q} = \mathbf{p}^{(s-2)}\mathbf{Q}^2 = \dots = \mathbf{p}^{(0)}\mathbf{Q}^s.$$

# Perron-Frobenius theorem

Let  $A$  be a square matrix, satisfying  $A \geq 0$  and  $A^k > 0$  for some  $k$ .

1. There exists a real eigenvalue  $\lambda_{PF} > 0$  with associated *positive* left/right eigenvectors.
2. For any other eigenvalue  $\lambda$  of  $A$ ,  $|\lambda| < |\lambda_{PF}|$
3.  $\lambda_{PF}$  has multiplicity 1 and corresponds to  $1 \times 1$  Jordan block.

# Transition matrix

Assume that our transition matrix satisfies  $Q^k > 0$  for some  $k$ . We know:

- ▶  $Q \geq 0$
- ▶ If  $\mathbb{1} = (1, \dots, 1)$ , then  $Q\mathbb{1}^T = \mathbb{1}^T$ , so 1 is an eigenvalue with right eigenvector  $\mathbb{1}^T$ .
- ▶ But the eigenvalues of  $Q$  satisfy  $|\lambda| \leq 1$  (Gershgorin circle theorem) .

Therefore  $\lambda_{PF} = 1$  and there exists a positive left eigenvector  $\pi$  for which

$$\pi Q = \pi \quad \text{and} \quad \pi \mathbb{1}^T = 1 \quad (\text{Why?})$$

We call such a  $\pi$  *the* stationary distribution.



## Stationary distributions

Because all other eigenvalues are bounded below 1, they die away, and

$$\lim_{s \rightarrow \infty} Q^s = \mathbb{1}^T \pi = \begin{pmatrix} -\pi- \\ \vdots \\ -\pi- \end{pmatrix}$$

On the other hand, even without the regularity assumption ( $Q^k > 0$ ), any limiting distribution is a stationary distribution. Take  $p$  an arbitrary limiting distribution:

$$\begin{aligned} \text{(assume)} \quad & \lim_{s \rightarrow \infty} Q^s = \mathbb{1}^T p \\ \text{(then)} \quad & \lim_{s \rightarrow \infty} Q^s Q = \mathbb{1}^T p Q \\ \text{(but)} \quad & \lim_{s \rightarrow \infty} Q^{s+1} = \mathbb{1}^T p \\ & \implies pQ = p \end{aligned}$$

# Law of large numbers

Consider a Markov chain with finite state space and regular transition matrix. If a function  $f(\cdot)$  is bounded on  $\mathcal{S}$ , then

$$\frac{1}{S} \sum_{s=0}^S f(\boldsymbol{\theta}^{(s)}) \xrightarrow{a.s.} \mathbb{E}_{\boldsymbol{\pi}} f(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} f(i) \pi_i.$$

This result holds irrespective of initial state  $\mathbf{p}^{(0)}$ .

# The punchline

- ▶ We construct Markov chains so that they have a specific stationary distribution  $\pi$  (e.g., the posterior).
- ▶ By simulating the Markovian dynamics, we may obtain an *empirical* estimate of  $\mathbb{E}_{\pi} f(\theta)$  .

# Detailed balance

Satisfying the *detailed balance* equations

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

is sufficient (assuming regularity, of course) for guaranteeing that  $\pi$  is the invariant distribution of the Markov chain:

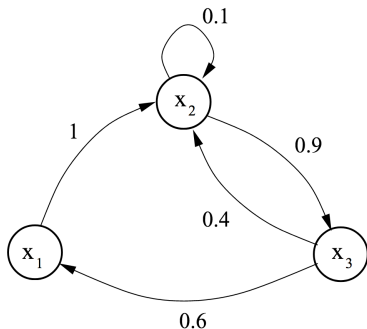
$$\sum_i \pi_i Q_{ij} = \sum_i \pi_j Q_{ji} = \pi_j \sum_i Q_{ji} = \pi_j$$

We say:

- ▶ The Markov chain is reversible with respect to  $\pi$  or
- ▶ the Markov chain satisfies detailed balance with respect to  $\pi$ .

## Two concepts

A chain is *irreducible* if for any two states  $i$  and  $j$ , there exists a  $k$  such that  $(Q^k)_{ij} > 0$ . Intuitively, this means the transition graph is connected.



Andrieu et al. 2003

The *period* of a state  $i$  is the *gcd* of the times at which it is possible to move from  $i$  to  $i$ . A Markov chain is *aperiodic* if the period of all states is 1.

# Existence and uniqueness of stationary distribution

Finite state space:

Irreducibility + Aperiodicity  $\iff$  Regular  $\iff$  Ergodic

Countable state space:

Irreducibility + Aperiodicity + Positive recurrence  $\iff$  Ergodic

A state is *positive recurrent* if the expected time to return is finite.

A chain is positive recurrent if all states are positive recurrent.

## Part 3. Discrete time, continuous space, time-homogeneous Markov chains

# Analogies: the Markov property

The Markov property

$$\Pr(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}, \dots, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(0)}) = \Pr(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)})$$

now becomes

$$\Pr(\boldsymbol{\theta}^{(s)} \in A | \boldsymbol{\theta}^{(s-1)}, \dots, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(0)}) = \Pr(\boldsymbol{\theta}^{(s)} \in A | \boldsymbol{\theta}^{(s-1)}).$$



## Analogies: transition kernel

The previous fact that

$$(p^{(s)})_j = (p^{(0)}Q^s)_j = \sum_{i_0, i_1, \dots, i_{s-2}, i_{s-1}} p_{i_0}^{(0)} Q_{i_0 i_1} \dots Q_{i_{s-2} i_{s-1}} Q_{i_{s-1} j}$$

becomes

$$\Pr(\boldsymbol{\theta}^{(s)} \in A) = \int_A p_s(\boldsymbol{\theta}^{(s)}) d\boldsymbol{\theta}^{(s)} = \int_A \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} q(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}) \dots q(\boldsymbol{\theta}^{(1)} | \boldsymbol{\theta}^{(0)}) p_0(\boldsymbol{\theta}^{(0)}) d\boldsymbol{\theta}^{(0)} \dots d\boldsymbol{\theta}^{(s-1)} d\boldsymbol{\theta}^{(s)},$$

i.e., we replace the transition matrix with the integral kernel

$$\int p_{s-1}(\boldsymbol{\theta}^{(s-1)}) q(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}) d\boldsymbol{\theta}^{(s-1)} = p_s(\boldsymbol{\theta}^{(s)}).$$

## Analogies: stationary distributions

The definition of a stationary distribution

$$\pi Q = \pi$$

becomes

$$\pi(\boldsymbol{\theta}^{(s)}) = \int q(\boldsymbol{\theta}^{(s)} | \boldsymbol{\theta}^{(s-1)}) \pi(\boldsymbol{\theta}^{(s-1)}) d\boldsymbol{\theta}^{(s-1)},$$

i.e.,  $\pi(\cdot)$  is an eigenfunction of the transition kernel with eigenvalue 1.

## Analogies: detailed balance

Detailed balance equations

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

becomes (a.s.)

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*).$$

If the chain satisfies detailed balance with respect to  $\pi(\cdot)$ , then

$$\int \pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)d\boldsymbol{\theta}^* = \int \pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})d\boldsymbol{\theta}^* = \pi(\boldsymbol{\theta}),$$

i.e.,  $\pi(\cdot)$  is a stationary distribution of the Markov chain.

## Useful concepts

- ▶ An MC is *p-irreducible* if there is a positive probability of reaching any set  $A$  for which  $\int_A p(\theta)d\theta > 0$ , regardless of initial state.
- ▶ A chain is *periodic* if it returns to any set  $A$  at regular intervals (*gcd* of return times  $> 1$ ). Otherwise it is *aperiodic*.

A sufficient condition for aperiodicity and  $p$ -irreducibility is that

$$\int_A q(\theta|\theta^{(0)})d\theta > 0, \forall \theta^{(0)} \quad \text{if} \quad \int_A p(\theta)d\theta > 0.$$

# Limiting distribution

If a chain has a stationary distribution  $\pi(\cdot)$  and is  $\pi$ -irreducible and aperiodic, then

1.  $\pi(\cdot)$  is the unique stationary distribution, and
2.  $\lim_{s \rightarrow \infty} \Pr(\boldsymbol{\theta}^{(s)} \in A | \boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}^*) = \int_A \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ ,

where we have asserted that the initial state has some value with probability 1.

# Existence and uniqueness of stationary distribution

Finite state space:

$$\text{Irreducibility} + \text{Aperiodicity} \iff \text{Regular} \iff \text{Ergodic}$$

Countable state space:

$$\text{Irreducibility} + \text{Aperiodicity} + \text{Positive recurrence} \iff \text{Ergodic}$$

Continuous state space:

$$\pi\text{-Irreducibility} + \text{Aperiodicity} + \text{Harris recurrence} \iff \text{Ergodic}$$

A state is *Harris recurrent* if for any starting value and any set  $A$  with  $\int_A \pi(\theta) d\theta > 0$ , the probability  $A$  is returned to infinitely often is 1.

# Consequences of ergodicity

For an ergodic chain with stationary distribution  $\pi(\cdot)$ ,

1.  $\lim_{s \rightarrow \infty} \Pr(\boldsymbol{\theta}^{(s)} \in A) = \int_A \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ , and

2.  $\frac{1}{S} \sum_{s=1}^S f(\boldsymbol{\theta}^{(s)}) \xrightarrow{a.s.} \mathbb{E}_{\pi} f(\boldsymbol{\theta})$ ,

provided the expectation is finite.

## In practice

Three things we can actually check:

1. Sufficient condition for  $\pi(\cdot)$  being a stationary distribution is reversibility / detailed balance:

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*).$$

2. Sufficient condition for aperiodicity and  $\pi$ -irreducibility is that

$$\int_A q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})d\boldsymbol{\theta} > 0, \forall \boldsymbol{\theta}^{(0)} \quad \text{if} \quad \int_A \pi(\boldsymbol{\theta})d\boldsymbol{\theta} > 0.$$

3. Sufficient condition for Harris recurrence is  $\pi$ -irreducibility and *absolute continuity* of  $q(\cdot|\boldsymbol{\theta}^*)$  wrt  $\pi(\cdot)$ :

$$\int_A \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = 0 \quad \implies \quad \int_A q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)d\boldsymbol{\theta}.$$



## Part 4. Classical MCMC

# Time for a 180°

So far:

$$\mathcal{S} + q(\cdot, \cdot) \implies \pi(\cdot)$$

Markov chain Monte Carlo:

$$\mathcal{S} + \pi(\cdot) \implies q(\cdot, \cdot)$$

## In practice

Three things we can actually check:

1. Sufficient condition for  $\pi(\cdot)$  being a stationary distribution is reversibility / detailed balance:

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*).$$

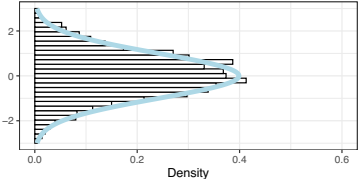
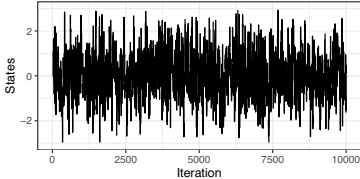
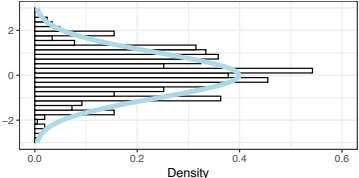
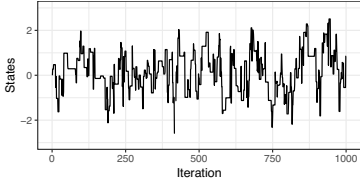
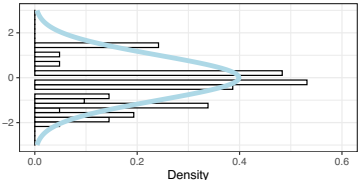
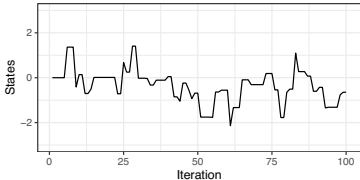
2. Sufficient condition for aperiodicity and  $\pi$ -irreducibility is that

$$\int_A q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})d\boldsymbol{\theta} > 0, \forall \boldsymbol{\theta}^{(0)} \quad \text{if} \quad \int_A \pi(\boldsymbol{\theta})d\boldsymbol{\theta} > 0.$$

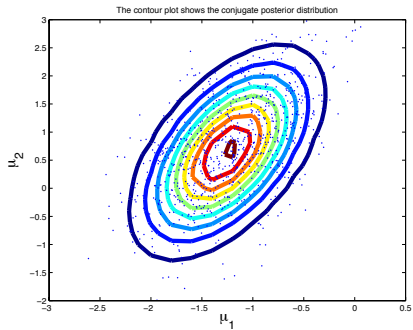
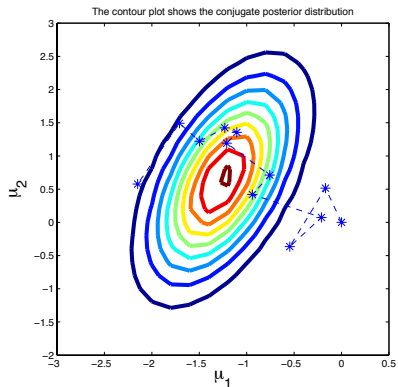
3. Sufficient condition for Harris recurrence is  $\pi$ -irreducibility and *absolute continuity* of  $q(\cdot|\boldsymbol{\theta}^*)$  wrt  $\pi(\cdot)$ :

$$\int_A \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = 0 \quad \implies \quad \int_A q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)d\boldsymbol{\theta}.$$

# Markov chain Monte Carlo



# Markov chain Monte Carlo



# The Metropolis algorithm

Our target stationary distribution is  $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|y) \propto p^*(\boldsymbol{\theta}|y)$ .

Inputs:

- ▶  $p^*(\boldsymbol{\theta}|y)$
- ▶ a *proposal distribution*  $h(\boldsymbol{\theta}^*|\boldsymbol{\theta})$  such that  $h(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = h(\boldsymbol{\theta}^*|\boldsymbol{\theta})$
- ▶  $\boldsymbol{\theta}^{(0)}$  (chosen or randomly generated however you want)

For  $s = 1, \dots, S$ ,

1. Generate  $\boldsymbol{\theta}^* \sim h(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s-1)})$  and  $U \sim \text{Uni}(0, 1)$
2. Compute

$$a \leftarrow 1 \wedge \frac{p^*(\boldsymbol{\theta}^*|y)}{p^*(\boldsymbol{\theta}^{(s-1)}|y)} = 1 \wedge \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(s-1)})}$$

3. IF  $U < a$ :  $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^*$ ;  
ELSE:  $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^{(s-1)}$

# The Metropolis algorithm

The Metropolis algorithm generates Markov chains that are reversible wrt the target distribution  $\pi(\boldsymbol{\theta})$ :

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

# The Metropolis algorithm

For unbounded targets (why?), the classic symmetric proposal is a Gaussian centered at the current state:

$$\boldsymbol{\theta}^* \sim h(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(s-1)}) \equiv N_D(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(s-1)}, \Sigma).$$



# The Metropolis algorithm

For unbounded targets (why?), the classic symmetric proposal is a Gaussian centered at the current state:

$$\boldsymbol{\theta}^* \sim h(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(s-1)}) \equiv N_D(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(s-1)}, \Sigma).$$

# Metropolis-Hastings

Our target stationary distribution is  $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|y) \propto p^*(\boldsymbol{\theta}|y)$ .

Inputs:

- ▶  $p^*(\boldsymbol{\theta}|y)$
- ▶ a not-necessarily-symmetric proposal distribution  $h(\boldsymbol{\theta}^*|\boldsymbol{\theta})$
- ▶  $\boldsymbol{\theta}^{(0)}$  (chosen or randomly generated however you want)

For  $s = 1, \dots, S$ ,

1. Generate  $\boldsymbol{\theta}^* \sim h(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s-1)})$  and  $U \sim \text{Uni}(0, 1)$
2. Compute

$$a \leftarrow 1 \wedge \frac{p^*(\boldsymbol{\theta}^*|y) h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p^*(\boldsymbol{\theta}^{(s-1)}|y) h(\boldsymbol{\theta}^*|\boldsymbol{\theta})} = 1 \wedge \frac{\pi(\boldsymbol{\theta}^*) h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(s-1)}) h(\boldsymbol{\theta}^*|\boldsymbol{\theta})}$$

3. IF  $U < a$ :  $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^*$ ;  
ELSE:  $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^{(s-1)}$

## Decomposing the parameter space

- ▶ Sometimes it is useful/easier to decompose the parameter space into several components.
- ▶ We want to use MH to sample from  $\pi(\boldsymbol{\theta}) = \pi(\theta_1, \dots, \theta_D)$ .
- ▶ Keep all but one component  $\theta_d$  fixed and use a univariate proposal to update  $\theta_d$ .

## Decomposing the parameter space

To update the  $d$ th component within global MCMC iteration  $s$  with state  $(\theta_1^{(s)}, \dots, \theta_{d-1}^{(s)}, \theta_d^{(s-1)}, \dots, \theta_D^{(s-1)})$ .

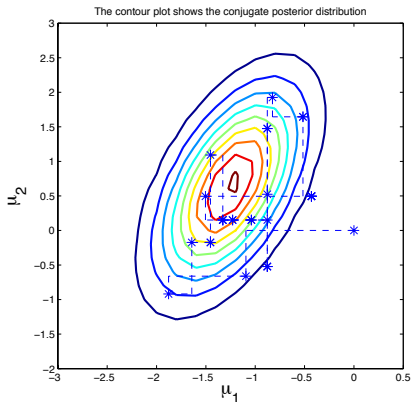
1. Propose  $\theta_d^* \sim h_d(\theta_d^* | \theta_1^{(s)}, \dots, \theta_{d-1}^{(s)}, \theta_d^{(s-1)}, \dots, \theta_D^{(s-1)})$   
 $\equiv h_d(\boldsymbol{\theta}^* | \boldsymbol{\theta})$

2. Accept with probability

$$1 \wedge \frac{\pi(\theta_1^{(s)}, \dots, \theta_{d-1}^{(s)}, \theta_d^*, \dots, \theta_D^{(s-1)}) h_d(\boldsymbol{\theta} | \boldsymbol{\theta}^*)}{\pi(\theta_1^{(s)}, \dots, \theta_{d-1}^{(s)}, \theta_d^{(s-1)}, \dots, \theta_D^{(s-1)}) h_d(\boldsymbol{\theta}^* | \boldsymbol{\theta})}$$

# Decomposing the parameter space

- ▶ We can decompose into blocks of components.
- ▶ We can use a random scan instead of sequential updates.
- ▶ If  $\pi(\theta)$  invariant to  $h_1$ ,  $h_2$ , then  $\pi(\theta)$  invariant to  $h_1 \circ h_2$ .



## Neat trick!

Suppose we divide  $\theta$  into two components:  $\theta = (\theta_1, \theta_2)$  and that

$$h_1(\theta_1|\theta_2) = \pi(\theta_1|\theta_2) = \pi(\theta)/\pi(\theta_2) = \pi(\theta) / \int \pi(\theta)d\theta_1$$

and analogous for  $h_2(\theta_2|\theta_1)$ . Then the MH acceptance criterion is  $\theta_1^{(s)}$

$$\begin{aligned} a &= 1 \wedge \frac{\pi(\theta_1^*, \theta_2^{(s-1)})}{\pi(\theta_1^{(s-1)}, \theta_2^{(s-1)})} \times \frac{\pi(\theta_1^{(s-1)}|\theta_2^{(s-1)})}{\pi(\theta_1^*|\theta_2^{(s-1)})} \\ &= 1 \wedge \frac{\pi(\theta_1^*, \theta_2^{(s-1)})}{\pi(\theta_1^{(s-1)}, \theta_2^{(s-1)})} \times \frac{\pi(\theta_1^{(s-1)}, \theta_2^{(s-1)})}{\pi(\theta_1^*, \theta_2^{(s-1)})} \times \frac{\pi(\theta_2^{(s-1)})}{\pi(\theta_2^{(s-1)})} = 1 \end{aligned}$$

and similar for  $\theta_2^{(s)}$ . Thus, we can avoid wasted compute time on rejected proposals.

Neat trick!

But when can we use it?

## Part 5. Introduction (?) to Bayesian inference



# Bayesian inference

We observe data  $y_1, \dots, y_N \stackrel{iid}{\sim} p(y_n|\boldsymbol{\theta})$  and assume  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ .  
Here,

- ▶  $p(y|\boldsymbol{\theta}) = \prod_{n=1}^N p(y_n|\boldsymbol{\theta})$  is the *likelihood*,
- ▶  $p(\boldsymbol{\theta})$  is the *prior*,

and the goal of Bayesian inference is to obtain the *posterior*

$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y)} = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

# Conjugate priors

- ▶ *Conjugacy* refers to the situation when the prior  $p(\theta)$  and posterior  $p(\theta|y)$  belong to the same distribution (albeit with “updated” parameters).
- ▶ When one combines a *conjugate* prior with a specific likelihood, one may obtain the posterior in closed form, no computations necessary!
- ▶ Unfortunately, conjugacy only works for a limited class of simple models.

# Exponential family distributions

- ▶ Exponential family distributions include the normal, beta, Bernoulli, gamma and Poisson distributions.
- ▶ If  $y$  follows an exponential family distribution, then

$$p(y|\theta) = h(y)g(\theta) \exp\left(\phi(\theta)^T s(y)\right).$$

- ▶ The joint distribution for independent  $y = (y_1, \dots, y_N)$  is

$$p(y|\theta) = \left(\prod_{n=1}^N h(y_n)\right) g^N(\theta) \exp\left(\phi(\theta)^T \sum_{n=1}^N s(y_n)\right).$$

- ▶  $\phi(\theta)$  is the *natural parameter* and  $t(y) = \sum_n s(y_n)$  is the *sufficient statistic*.

## Conjugate priors

Again, our likelihood is

$$p(y|\boldsymbol{\theta}) \propto g^N(\boldsymbol{\theta}) \exp\left(\phi(\boldsymbol{\theta})^T t(y)\right),$$

and we specify  $\boldsymbol{\theta}$  follows an exponential family distribution with prior

$$p(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})^\eta \exp\left(\phi(\boldsymbol{\theta})^T \nu\right).$$

It follows that

$$p(\boldsymbol{\theta}|y) \propto g^{N+\eta}(\boldsymbol{\theta}) \exp\left(\phi(\boldsymbol{\theta})^T (t(y) + \nu)\right).$$

## Beta-binomial model

$$p(y|\theta, N) \propto \theta^y (1 - \theta)^{N-y} \propto (1 - \theta)^N \exp\left(y \log\left(\frac{\theta}{1 - \theta}\right)\right)$$

$$\implies g(\theta) = 1 - \theta \quad \text{and} \quad \phi(\theta) = \log\left(\frac{\theta}{1 - \theta}\right)$$

$$\implies p(\theta) \propto (1 - \theta)^\eta \exp\left(\nu \log\left(\frac{\theta}{1 - \theta}\right)\right) \propto (1 - \theta)^{\eta - \nu} \theta^\nu$$

$$\implies p(\theta) \equiv \text{beta}(\alpha = \nu + 1, \beta = \eta - \nu + 1)$$

$$\implies p(\theta|y) \propto (1 - \theta)^{(\eta - \nu + N - y)} \theta^{\nu + y}$$

$$\implies p(\theta|y) \equiv \text{beta}(\alpha + y, \beta + N - y)$$

$$\implies \mathbb{E}(\theta|y) = (\alpha + y) / (\alpha + \beta + N)$$

## Univariate normal, known variance

$$p(y|\theta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_n (y_n - \theta)^2\right) \propto \exp\left(-\frac{N\theta^2}{2\sigma^2} + \frac{\theta}{\sigma^2} \sum_n y_n\right)$$

$$\implies p(\theta) \propto \exp\left(-\frac{\theta^2}{2\tau_0^2} + \frac{\mu_0\theta}{\tau_0^2}\right) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right)$$

$$\begin{aligned}\implies p(\theta|y, \sigma^2) &\propto \exp\left(-\frac{\theta^2}{2\tau_0^2} + \frac{\mu_0\theta}{\tau_0^2}\right) \exp\left(-\frac{N\theta^2}{2\sigma^2} + \frac{\theta}{\sigma^2} \sum_n y_n\right) \\ &\propto \exp\left(-\frac{1}{2} \left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right) \theta^2 + \left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right) \theta\right) \\ &\equiv \mathbf{N}\left(\left(\frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2}\right) \left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}, \left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1}\right)\end{aligned}$$

## Univariate normal, known mean

$$p(y|\theta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_n (y_n - \theta)^2\right)$$

$$\implies p(\sigma^2) \propto (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right) \equiv \Gamma^{-1}(\alpha, \beta)$$

$$\begin{aligned} \implies p(\sigma^2|y, \theta) &\propto (\sigma^2)^{-\alpha-N/2-1} \exp\left(-\frac{\beta}{\sigma^2} + \frac{\sum_n (y_n - \theta)^2}{2\sigma^2}\right) \\ &\equiv \Gamma^{-1}\left(\alpha + \frac{N}{2}, \beta + \frac{\sum_n (y_n - \theta)^2}{2}\right) \end{aligned}$$

## Limitations to conjugacy

- ▶ We rarely know the variance but not the mean (and vice-versa).
- ▶ We don't have the joint posterior for both mean and variance in closed form.
- ▶ All we know is the conditional posteriors for either parameter.
- ▶ It turns out, this kind of situation is rather common for Bayesian hierarchical models that arise out of pieced together exponential family distributions.



## Part 6. Classical MCMC (again)

## Neat trick!

Suppose we divide  $\theta$  into two components:  $\theta = (\theta_1, \theta_2)$  and that

$$h_1(\theta_1|\theta_2) = \pi(\theta_1|\theta_2) = \pi(\theta)/\pi(\theta_2) = \pi(\theta) / \int \pi(\theta)d\theta_1$$

and analogous for  $h_2(\theta_2|\theta_1)$ . Then the MH acceptance criterion is  $\theta_1^{(s)}$

$$\begin{aligned} a &= 1 \wedge \frac{\pi(\theta_1^*, \theta_2^{(s-1)})}{\pi(\theta_1^{(s-1)}, \theta_2^{(s-1)})} \times \frac{\pi(\theta_1^{(s-1)}|\theta_2^{(s-1)})}{\pi(\theta_1^*|\theta_2^{(s-1)})} \\ &= 1 \wedge \frac{\pi(\theta_1^*, \theta_2^{(s-1)})}{\pi(\theta_1^{(s-1)}, \theta_2^{(s-1)})} \times \frac{\pi(\theta_1^{(s-1)}, \theta_2^{(s-1)})}{\pi(\theta_1^*, \theta_2^{(s-1)})} \times \frac{\pi(\theta_2^{(s-1)})}{\pi(\theta_2^{(s-1)})} = 1 \end{aligned}$$

and similar for  $\theta_2^{(s)}$ . Thus, we can avoid wasted compute time on rejected proposals.

Neat trick!

But when can we use it?

## A Gibbs sampler

We assume our data  $y = (y_1, \dots, y_N) \stackrel{iid}{\sim} N(\theta, \sigma^2)$  and priors

$$\theta \sim N(\mu_0, \tau_0^2) \quad \text{and} \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

We wish to generate samples from  $p(\theta, \sigma^2 | y)$ . Initialize  $\theta^{(0)}$  and  $\sigma^{(0)}$ . For  $s = 1, \dots, S$ ,

1. Draw from  $p(\theta | y, \sigma^2)$  with  $\sigma^2 = \sigma^{2(s-1)}$ :

$$\theta^{(s)} \sim N \left( \left( \frac{\mu_0}{\tau_0^2} + \frac{\sum_n y_n}{\sigma^2} \right) \left( \frac{1}{\tau_0^2} + \frac{N}{\sigma^2} \right)^{-1}, \left( \frac{1}{\tau_0^2} + \frac{N}{\sigma^2} \right)^{-1} \right).$$

2. Draw from  $p(\sigma^2 | y, \theta)$  with  $\theta = \theta^{(s)}$ :

$$\sigma^{2(s)} \sim \Gamma^{-1} \left( \alpha + \frac{N}{2}, \beta + \frac{\sum_n (y_n - \theta)^2}{2} \right)$$

No need for the accept/reject step!

## Another Gibbs sampler

We assume our data  $y_n \stackrel{ind}{\sim} N(\theta_n, \sigma^2)$ ,  $n = 1, \dots, N$ ,

$$\theta_n \stackrel{iid}{\sim} N(\theta_0, \tau_0^2) \quad \text{and} \quad \theta_0 \sim N(0, 10).$$

We wish to sample from  $p(\theta_0, \theta_1, \dots, \theta_N | y, \sigma^2, \tau^2)$ . After initialization, for  $s = 1, \dots, S$ :

1. Draw from  $p(\theta_0 | y, \tau^2, \theta_1^{(s-1)}, \dots, \theta_N^{(s-1)})$ :

$$\theta_0^{(s)} \sim N \left( \left( \frac{\sum_n \theta_n^{(s-1)}}{\tau_0^2} \right) \left( \frac{N}{\tau_0^2} + \frac{1}{10} \right)^{-1}, \left( \frac{N}{\tau_0^2} + \frac{1}{10} \right)^{-1} \right)$$

2. For  $n = 1, \dots, N$ , draw from  $p(\theta_n | y, \sigma^2, \tau^2, \theta_0^{(s)})$ :

$$\theta_1^{(s)} \sim N \left( \left( \frac{\theta_0^{(s)}}{\tau_0^2} + \frac{y_n}{\sigma^2} \right) \left( \frac{1}{\tau_0^2} + \frac{1}{\sigma^2} \right)^{-1}, \left( \frac{1}{\tau_0^2} + \frac{1}{\sigma^2} \right)^{-1} \right).$$

# Pros and cons of Gibbs sampling

## Pros:

- ▶ No wasted compute time on rejected proposals.
- ▶ For big data, factorization helps
  1. data storage
  2. parallel computing.

## Cons:

- ▶ You're only as strong as your weakest link. (But isn't this always true?)
- ▶ Coding by hand can be time intensive. (But isn't there software for that?)
- ▶ Conditional posteriors aren't always known. (But isn't there Metropolis-within-Gibbs for that?)