

# MARKOV CHAIN MONTE CARLO (MCMC) METHODS

---

<sup>0</sup>These notes utilize a few sources: some insights are taken from Profs. Vardeman's and Carriquiry's lecture notes, some from a great book on Monte Carlo strategies in scientific computing by J.S. Liu.

# Markov Chains: Basic Theory

**Definition 1.** A (discrete time/discrete state space) *Markov chain (MC)* is a sequence of random quantities  $\{X_k\}$ , each taking values in a (finite or) countable set  $\mathcal{X}$ , with the property that

$$\begin{aligned} P\{X_n = x_n \mid X_1 = x_1, \dots, X_{n-1} = x_{n-1}\} \\ = P\{X_n = x_n \mid X_{n-1} = x_{n-1}\}. \end{aligned}$$

**Definition 2.** A Markov Chain is stationary if

$$P\{X_n = x \mid X_{n-1} = x'\}$$

is independent of  $n$ .

Without loss of generality, we will henceforth name the elements of  $\mathcal{X}$  with the integers  $1, 2, 3, \dots$  and call them “states.”

**Definition 3.** With

$$p_{ij} \triangleq P\{X_n = j \mid X_{n-1} = i\}$$

the square matrix

$$P \triangleq \{p_{ij}\}$$

is called the **transition matrix** for a stationary Markov Chain and the  $p_{ij}$  are called **transition probabilities**.

Note that a transition matrix has nonnegative entries and its rows sum to 1. Such matrices are called **stochastic matrices**.

More notation for a stationary MC: Define

$$p_{ij}^{(k)} \triangleq P\{X_{n+k} = j \mid X_n = i\}$$

and

$$f_{ij}^{(k)} \triangleq P\{X_{n+k} = j, X_{n+k-1} \neq j, \dots, X_{n+1} \neq j, \mid X_n = i\}.$$

These are respectively the probabilities of moving from  $i$  to  $j$  in  $k$  steps and first moving from  $i$  to  $j$  in  $k$  steps.

**Definition 4.** We say that a MC is **irreducible** if, for each  $i$  and  $j$ , there exists a  $k$  (possibly depending upon  $i$  and  $j$ ) such that

$$p_{ij}^{(k)} > 0 \quad \text{for finite } k.$$

Or, in words, a chain is irreducible if it is possible to eventually get from any state  $i$  to any other state  $j$  in finite number of steps.

**Example:** The chain with transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{bmatrix}$$

is irreducible. But, the chain with transition matrix

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is reducible.

Consider this block structure for the transition matrix:

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}, \quad P_1, P_2 \text{ are } 2 \times 2 \text{ matrices}$$

where the overall chain is reducible, but its pieces (sub-chains)  $P_1$  and  $P_2$  could be irreducible.

**Definition 5.** We say that the  $i$ th state of a MC is *transient* if

$$\sum_{k=0}^{\infty} f_{ii}^{(k)} < 1$$

and say that this state is *persistent* (or *recurrent* which is the term used as well to describe this phenomenon) if

$$\sum_{k=0}^{\infty} f_{ii}^{(k)} = 1.$$

(Note that  $f_{ii}^{(0)} = 0$ .) A chain is called persistent if all of its states are persistent. In words, a state is transient if once in

it, there is some possibility that the chain will never return. A state is persistent (recurrent) if once in it, the chain will with certainty be in it again.

**Definition 6.** We say that state  $i$  of a MC has *period*  $t$  if  $p_{ii} = 0$  unless  $k = \nu t$  (where  $\nu$  is an integer, i.e.  $k$  is an integer multiple of  $t$ ) and  $t$  is the largest integer with this property. The state is aperiodic if no such  $t > 1$  exists. An MC is called aperiodic if all of its states are aperiodic.

Many sources (including Chapter 15 of the 3rd Edition of Feller vol. 1) present a number of useful simple results about MCs. Among them are the following.

**Theorem 1.** All states of an irreducible MC are of the same type with regard to persistence (recurrence) and periodicity.

**Theorem 2.** A finite-state-space irreducible MC is persistent (recurrent).

**Theorem 3.** A state  $i$  is persistent (recurrent) iff

$$\sum_{k=0}^{\infty} p_{ii}^{(k)} = \infty.$$

This is an important theorem!

**Proof.** Interestingly, we can prove this theorem using the  $z$  transform (which, in this context, is known as the probability generating function).

Note that  $p_{ii}^{(0)} = 1$  and, for  $n \geq 1$ , we have

$$\begin{aligned}
 p_{ii}^{(n)} &= P\{X_n = i \mid X_0 = i\} \\
 &= \underbrace{P\{X_1 = i \mid X_0 = i\}}_{f_{ii}^{(1)}} \cdot \underbrace{P\{X_{n-1} = i \mid X_0 = i\}}_{p_{ii}^{(n-1)}} \\
 &\quad + \sum_{k=2}^n \underbrace{P\{X_k = i, X_{k-1} \neq i, \dots, X_1 \neq i \mid X_0 = i\}}_{f_{ii}^{(k)}} \\
 &\quad \cdot \underbrace{P\{X_{n-k} = i \mid X_0 = i\}}_{p_{ii}^{(n-k)}}.
 \end{aligned}$$

Note also that  $f_{ii}^{(0)} = 0$ .

Combining the above facts, we write a general formula that holds for all  $n \geq 0$ :

$$p_{ii}^{(n)} = \delta_{n,0} + \sum_{k=0}^n f_{ii}^{(k)} p_{ii}^{(n-k)}.$$

Now, take the  $z$  transform of the above expression:

$$P_{ii}(z) = 1 + F_{ii}(z) P_{ii}(z)$$

and, consequently,

$$P_{ii}(z) = \frac{1}{1 - F_{ii}(z)}.$$

State  $i$  is persistent (recurrent) if  $\sum_{k=0}^{\infty} f_{ii}^{(k)} = 1$ , which is equivalent to

$$F_{ii}(z) \Big|_{z=1-} = 1$$

which further implies

$$P_{ii}(z) \Big|_{z=1-} = +\infty.$$

Conversely, state  $i$  is transient if  $\sum_{k=0}^{\infty} f_{ii}^{(k)} < 1$ , which is equivalent to

$$F_{ii}(z) \Big|_{z=1-} < 1$$

which further implies

$$P_{ii}(z) \Big|_{z=1-} < +\infty.$$

□

**Note:** We can interpret  $P_{ii}(z) \Big|_{z=1-}$  as the expected number of visits to state  $i$ , starting from state  $i$ . Indeed, this expected

number of visits is

$$\mathbb{E} \left[ \sum_{k=0}^{\infty} \delta_{X_k, i} \mid X_0 = i \right] = \sum_{k=0}^{\infty} p_{ii}^{(k)} = P_{ii}(z) \Big|_{z=1-}.$$

Clearly, if state  $i$  is persistent (recurrent), we will keep visiting it over and over again, and the expected number of visits is infinite!

**Definition 7.** We define the time of first “hitting” state  $i_0$  as

$$T_{i_0} = \min\{n : n \geq 1, X_n = i_0\}.$$

**Comments:** Clearly, state  $i_0$  is persistent (recurrent) if

$$P\{T_{i_0} < \infty \mid X_0 = i_0\} = 1.$$

**Definition 8.** A persistent (recurrent) state  $i_0$  is *positive persistent (recurrent)* if

$$\underbrace{\mathbb{E}[T_{i_0} \mid X_0 = i_0]}_{\text{mean recurrence time}} = \sum_{n=0}^{\infty} n f_{i_0 i_0}^{(n)} < \infty$$

and it is *null recurrent* if

$$\mathbb{E}[T_{i_0} \mid X_0 = i_0] = \infty.$$



So, the expected number of steps that it takes for our chain to revisit a positive-recurrent state is finite.

**Lemma 1.** *Suppose that state  $i$  persistent (recurrent). Then, this state is positive persistent (recurrent) if and only if*

$$\pi_i = \lim_{z \rightarrow 1} (1 - z)P_{ii}(z) > 0$$

and then

$$\pi_i = \frac{1}{\sum_{n=0}^{\infty} n f_{ii}^{(n)}} = \frac{1}{\mathbb{E}[T_i | X_0 = i]}.$$

**Proof.** From the proof of Theorem 3, we know

$$P_{ii}(z) = \frac{1}{1 - F_{ii}(z)}$$

and, therefore,

$$\frac{1 - F_{ii}(z)}{1 - z} = \frac{1}{(1 - z)P_{ii}(z)}.$$

Now

$$\lim_{z \rightarrow 1} \frac{1 - F_{ii}(z)}{1 - z} = F'_{ii}(z) \Big|_{z=1} = \sum_{n=0}^{\infty} n f_{ii}^{(n)} = \mathbb{E}[T_i | X_0 = i].$$

□

**Note:** This lemma almost proves the convergence of averages of transition probabilities, namely

$$\frac{1}{n+1} \sum_{k=0}^n p_{ii}^{(k)} \rightarrow \frac{1}{\mathbb{E}[T_i | X_0 = i]} \quad \text{as } n \rightarrow +\infty$$

Consider

$$(1 - z^{-1})P_{ii}(z) = \frac{\sum_{k=0}^{\infty} p_{ii}^{(k)} z^{-k}}{\sum_{k=0}^{\infty} z^{-k}} = \lim_{n \rightarrow \infty} \frac{\sum_{k=0}^n p_{ii}^{(k)} z^{-k}}{\sum_{k=0}^n z^{-k}}$$

for any  $z$  such that  $|z| \geq 1$ . If we could take the limit as  $z \rightarrow 1$  under the summations, we would get, using the above lemma:

$$\pi_i = \lim_{n \rightarrow \infty} \frac{1}{n+1} \cdot \sum_{k=0}^n p_{ii}^{(k)} = \frac{1}{\mathbb{E}[T_i | X_0 = i]}.$$

(This proof is not complete, since we did not verify that we can take the limit as  $z \rightarrow 1$  under the summations but the result holds.) The following theorems formalize the key results, part of whose proof we have hinted above.

**Theorem 4.** *Suppose that an MC is irreducible, aperiodic and persistent (recurrent). Suppose further that, for each state  $i$ , the mean recurrence time is finite, i.e.*

$$\mathbb{E}[T_{i_0} | X_0 = i_0] = \sum_{k=0}^{\infty} k f_{ii}^{(k)} < \infty$$

implying that the entire chain is positive recurrent! Then, an invariant/stationary distribution for the MC exists, i.e. there exist  $\pi_j$  with  $\pi_j > 0$  and  $\sum_j \pi_j = 1$  such that

$$\pi_j = \sum_i \pi_i p_{ij} \quad \text{(equation of full balance)}$$

or, in the matrix notation

$$\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T, \quad \boldsymbol{\pi} = [\pi_1, \pi_2, \dots]^T.$$

In words, if the chain is started with initial distribution  $\{\pi_j\}$  (of the states), then, after one transition, it is in states  $1, 2, 3, \dots$  with probabilities  $\{\pi_1, \pi_2, \pi_3, \dots\}$ . Further, this distribution  $\{\pi_j\}$  satisfies

$$\pi_j = \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=0}^n p_{ij}^{(k)}, \quad \forall i$$

and

$$\pi_j = \frac{1}{\sum_{k=0}^{\infty} k f_{jj}^{(k)}} = \frac{1}{\mathbb{E}[T_j | X_0 = j]}.$$

## Comments:

(A little) verification that, for stationary distributions, after one transition, the chain is in states  $1, 2, 3, \dots$  with probabilities

$\{\pi_1, \pi_2, \pi_3, \dots\}$ :

$$\begin{aligned} P\{X_1 = j\} &= \sum_i P\{X_1 = j, X_0 = i\} \\ &= \sum_i \underbrace{P\{X_1 = j | X_0 = i\}}_{p_{ij}} \underbrace{P\{X_0 = i\}}_{\pi_i} = \pi_j! \end{aligned}$$

Here is how we can compute:

$$E[g(X_1) | X_0 = i] = \sum_j p_{ij} g(j)$$

where  $g(\cdot)$  is a real function. Or, in matrix form (for a finite-state-space MC):

$$\begin{bmatrix} E[g(X_1) | X_0 = 1] \\ E[g(X_1) | X_0 = 2] \\ \vdots \\ E[g(X_1) | X_0 = N] \end{bmatrix} = P \begin{bmatrix} g(1) \\ g(2) \\ \vdots \\ g(N) \end{bmatrix}.$$

Converse to Theorem 4:

**Theorem 5.** *An irreducible, aperiodic MC for which there exists  $\{\pi_j\}$  with  $\pi_j > 0$  and  $\sum \pi_j = 1$  such that*

$$\pi_j = \lim_{k \rightarrow \infty} p_{ij}^{(k)}, \quad \forall i$$

*must be persistent (recurrent) with*

$$\pi_j = \frac{1}{\sum_{k=1}^{\infty} k f_{jj}^{(k)}}.$$

There is an important “ergodic” result that guarantees that “time averages” have the right limits:

**Theorem 6.** *Under the hypotheses of Theorem 4, if  $g(\cdot)$  is a real-valued function such that*

$$\sum_j |g(j)| \pi_j < \infty$$

*then, for any  $j$ , if  $X_0 = j$*

$$\frac{1}{n+1} \sum_{k=0}^n g(X_k) \rightarrow \frac{1}{n+1} \sum_j g(j) \pi_j \quad \text{as } n \rightarrow +\infty.$$

Note that the choice of  $g(\cdot)$  as an indicator (Kronecker delta) provides approximations for stationary probabilities:

$$\frac{1}{n+1} \sum_{k=0}^n \delta_{X_k, i} \rightarrow \pi_i \quad \text{as } n \rightarrow +\infty.$$

With this background, the basic idea of MCMC is the following. If we wish to simulate from a distribution  $\{\pi_j\}$  or approximate properties of the distribution that can be expressed in terms of some function  $g$ , we find a convenient MC  $\{X_k\}$  whose invariant distribution is  $\{\pi_j\}$ . From a starting state  $X_0 = i_0$ , one uses  $P$  to simulate  $X_1$ . Using the realization  $X_1 = x_1$  and  $P$ , one simulates  $X_2$  etc. We apply Theorem 6 to approximate the quantity of interest. Actually, it is common practice to use a “burn in” (i.e. discard a certain number of initial samples  $X_0, X_1, \dots, X_{\text{burnin}-1}$ ) before starting the kind of time average indicated in Theorem 6.

# Markov Chain Simulations

**Basic Idea:** Suppose the sampling from  $p(\boldsymbol{\theta}|\mathbf{x})$  is hard, but that we can somehow generate a Markov chain  $\{\boldsymbol{\theta}(t), t \in T\}$  with stationary distribution  $p(\boldsymbol{\theta}|\mathbf{x})$ .

**Note:** Here, we know the stationary distribution and seek transitions

$$\mathcal{P}(\boldsymbol{\theta}^{(t+1)} | \boldsymbol{\theta}^{(t)})$$

that will take us to this stationary distribution.

We will start from some initial guess  $\boldsymbol{\theta}^{(0)}$  and let the chain run for  $n$  steps (where  $n$  is large), i.e. until it reaches its stationary distribution. Upon convergence, all additional steps are draws from the stationary distribution  $p(\boldsymbol{\theta}|\mathbf{x})$ .

All MCMC methods are based on the same idea — the difference is just in how the transitions in the Markov chain are created.

# Gibbs Sampler (Again) in the Bayesian Context

**Idea:** Cycle through all possible full conditional posterior distributions. For example, suppose that  $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^T$  and that the target distribution is  $p(\boldsymbol{\theta}|\mathbf{x})$ . Then, the steps of the Gibbs sampler are

(a) Start with a guess  $\boldsymbol{\theta}^{(0)} = [\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}]^T$ .

(b) Draw  $\theta_1^{(1)}$  from

$$p(\theta_1 | \theta_2 = \theta_2^{(0)}, \theta_3 = \theta_3^{(0)}, \mathbf{x}).$$

(c) Draw  $\theta_2^{(1)}$  from

$$p(\theta_2 | \theta_1 = \theta_1^{(1)}, \theta_3 = \theta_3^{(0)}, \mathbf{x}).$$

(d) Draw  $\theta_3^{(1)}$  from

$$p(\theta_3 | \theta_1 = \theta_1^{(1)}, \theta_2 = \theta_2^{(1)}, \mathbf{x}).$$



The above steps complete one cycle of the Gibbs sampler.

Repeat the above steps  $n$  times (i.e. until the chain has converged) and, upon convergence, the draws  $\boldsymbol{\theta}^{(n+1)}, \boldsymbol{\theta}^{(n+2)}, \dots$  are samples from the stationary distribution  $p(\boldsymbol{\theta} | \boldsymbol{x})$ . Here,  $0, 1, \dots, n$  is called the “burn-in” period.

**Note:** We can update the  $\theta$ s one at a time (as above) or in blocks.

# A (Toy) Discrete Example: How Gibbs Can Fail

Suppose  $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$  and that the pmf  $p(\boldsymbol{\theta} | \boldsymbol{x})$  is described by the following table:

$\theta_2 \backslash \theta_1$	1	2	3	4
4	0	0	+	+
3	0	0	+	+
2	+	+	0	0
1	+	+	0	0

Here, the + signs indicate some positive probabilities that need to sum to 1. Start with  $\boldsymbol{\theta}^{(0)} = [\theta_1^{(0)}, \theta_2^{(0)}]^T$ . If we use Gibbs sample and start with  $\boldsymbol{\theta}^{(0)}$  in the upper-right corner, we will never escape the upper-right corner (of the probability space)! Similarly, if we start in the lower-left corner, we will never leave that part of the probability space. In MC terminology, our chain is reducible!

Regardless of how large burn-in period  $0, 1, \dots, n$  we choose, no sequence  $\boldsymbol{\theta}^{(n+1)}, \boldsymbol{\theta}^{(n+2)}, \dots$  can possibly “look like” coming from the distribution described by the above table!

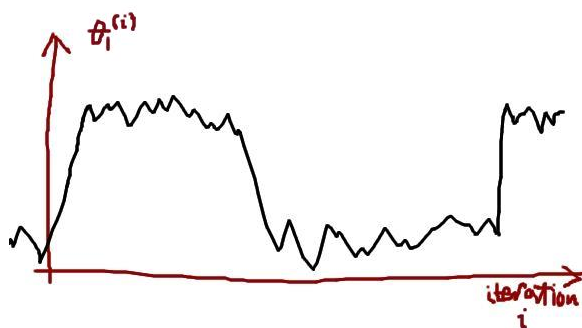
**Morale of this story:** We should not start from one place only! If we start from only one place, we will never detect the above problem.

If we are lucky and smart enough to run several chains (using several starting values), we may detect a problem if we see different chains giving different results.

Here is another scenario that may happen in practice:

$\theta_2 \backslash \theta_1$	1	2	3	4
4	0	0	+	+
3	0	0	+	+
2	+	+	$\epsilon$	0
1	+	+	0	0

where  $\epsilon$  is very small. In this case, Gibbs will produce time plots like



We would find that correlations between, say  $\theta_1^{(i)}$  and  $\theta_1^{(i+1)}$  are very large in this case. To fairly represent  $p(\boldsymbol{\theta}|\mathbf{x})$ , we would need a very large number of iterations. In this case, the problem is “poor mixing,” caused here by relatively isolated islands of probability.

Clearly, poor mixing may occur due to

- difficult pdf/pmf to sample from (as in the above example)  
or
- poor sampling algorithm (perhaps not tuned well)

or both.

## Why Gibbs Might Work

For simplicity, consider a 3-dimensional case with discrete probability space for the random variables  $\alpha$ ,  $\beta$ , and  $\gamma$  making

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}.$$

We wish to sample from  $p(\boldsymbol{\theta}) = p(\alpha, \beta, \gamma)$ . Recall the Gibbs sampler: start with some  $\boldsymbol{\theta}^{(0)} = \begin{bmatrix} \alpha^{(0)} \\ \beta^{(0)} \\ \gamma^{(0)} \end{bmatrix}$  and draw new samples by cycling as follows:

$$\boldsymbol{\theta}^{(1)} = \begin{bmatrix} \alpha^{(1)} \\ \beta^{(0)} \\ \gamma^{(0)} \end{bmatrix} \quad \text{where} \quad \alpha^{(1)} \sim p_{\alpha|\beta,\gamma}(\cdot | \beta^{(0)}, \gamma^{(0)})$$

$$\boldsymbol{\theta}^{(2)} = \begin{bmatrix} \alpha^{(1)} \\ \beta^{(1)} \\ \gamma^{(0)} \end{bmatrix} \quad \text{where} \quad \beta^{(1)} \sim p_{\beta|\alpha,\gamma}(\cdot | \alpha^{(1)}, \gamma^{(0)})$$

$$\boldsymbol{\theta}^{(3)} = \begin{bmatrix} \alpha^{(1)} \\ \beta^{(1)} \\ \gamma^{(1)} \end{bmatrix} \quad \text{where} \quad \gamma^{(1)} \sim p_{\gamma|\alpha,\beta}(\cdot | \alpha^{(1)}, \beta^{(1)})$$

$$\boldsymbol{\theta}^{(4)} = \begin{bmatrix} \alpha^{(2)} \\ \beta^{(1)} \\ \gamma^{(1)} \end{bmatrix} \quad \text{where} \quad \alpha^{(2)} \sim p_{\alpha|\beta,\gamma}(\cdot | \beta^{(1)}, \gamma^{(1)})$$

etc.

For example, for an “ $\alpha$  substitution,” we have:

$$\mathcal{P}\left(\boldsymbol{\theta}^{(t)} = \begin{bmatrix} \alpha' \\ \beta \\ \gamma \end{bmatrix} \mid \boldsymbol{\theta}^{(t-1)} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}\right) = \frac{p(\alpha', \beta, \gamma)}{\sum_a p(a, \beta, \gamma)}.$$

Now, consider the case where  $\boldsymbol{\theta}^{(t-1)}$  is coming from  $p(\alpha, \beta, \gamma)$ . Then, using the total probability theorem, we can compute the pmf of  $\boldsymbol{\theta}^{(t)}$ :

$$\begin{aligned} \mathcal{P}\left(\boldsymbol{\theta}^{(t)} = \begin{bmatrix} \alpha' \\ \beta \\ \gamma \end{bmatrix}\right) &= \sum_{\alpha} \mathcal{P}\left(\boldsymbol{\theta}^{(t)} = \begin{bmatrix} \alpha' \\ \beta \\ \gamma \end{bmatrix} \mid \boldsymbol{\theta}^{(t-1)} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}\right) \\ &\quad \cdot p(\alpha, \beta, \gamma) \\ &= \sum_{\alpha} \frac{p(\alpha', \beta, \gamma)}{\underbrace{\sum_a p(a, \beta, \gamma)}} \cdot p(\alpha, \beta, \gamma) \\ &\quad \text{does not depend on } \alpha \\ &= \frac{p(\alpha', \beta, \gamma)}{\sum_a p(a, \beta, \gamma)} \cdot \sum_{\alpha} p(\alpha, \beta, \gamma) = p(\alpha', \beta, \gamma). \end{aligned}$$

So, if  $\boldsymbol{\theta}^{(t-1)}$  comes from  $p(\cdot)$ , then  $\boldsymbol{\theta}^{(t)}$  comes from  $p(\cdot)$  as

well (equation of full balance)! The same kind of reasoning shows that the equation of full balance also holds for  $\beta$  and  $\gamma$  substitutions.

# General Theoretical Conditions for MCMC Methods to Work

See the MC theory at this beginning of this handout. In words, the conditions of the theorems from the MC theory part are:

- irreducibility  $\equiv$  no isolated islands of probability;
- aperiodicity  $\equiv$  no cyclic structure where we can get back to a state in some set number of transitions ( $> 1$ ) or a multiple thereof;
- persistence (recurrence)  $\equiv$  we are guaranteed to get back to any state that we leave.



# A Gibbs Sampling Example

(from STAT 544 notes by Prof. Carriquiry)

Suppose we have observations  $x_1, x_2, \dots, x_n$  following two Poisson distributions:

$$x_i \sim \text{Poisson}(\lambda), \quad \text{for } i = 1, 2, \dots, m$$

$$x_i \sim \text{Poisson}(\phi), \quad \text{for } i = m + 1, m + 2, \dots, n$$

where the change point  $m$  is unknown. The unknown parameters are  $\lambda, \phi$ , and  $m$ . Define  $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ .

We use a Bayesian approach to estimate these parameters. Choose the following priors:

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$

$$\phi \sim \text{Gamma}(\gamma, \delta)$$

$$m = \text{uniform}(1, 2, \dots, n).$$

Then, the joint posterior distribution is

$$\begin{aligned}
 p(\lambda, \phi, m | \mathbf{x}) & \propto \underbrace{\lambda^{\alpha-1} \exp(-\lambda\beta) \cdot \phi^{\gamma-1} \exp(-\phi\delta) \cdot i_{\{1,2,\dots,n\}}(m)}_{\text{prior}} \\
 & \underbrace{\prod_{i=1}^m \exp(-\lambda) \lambda^{x_i} \cdot \prod_{j=m+1}^n \exp(-\phi) \phi^{x_j}}_{\text{likelihood}} \\
 & \propto \lambda^{x_{1,\star}(m)+\alpha-1} \exp[-\lambda \cdot (m + \beta)] \\
 & \quad \cdot \phi^{x_{2,\star}(m)+\gamma-1} \exp[-\phi \cdot (n - m + \delta)] \cdot i_{\{1,2,\dots,n\}}(m) \\
 & \triangleq f(m, \lambda, \phi, \mathbf{x})
 \end{aligned}$$

where  $x_{1,\star}(m) = \sum_{i=1}^m x_i$  and  $x_{2,\star}(m) = \sum_{i=m+1}^n x_i$ .

**Note:** If we knew  $m$ , the problem would be trivial to solve, since we chose conjugate priors for  $\lambda$  and  $\phi$ . When we do not know  $m$ , we can use the Gibbs sampler.

We need to find full conditionals. Select pieces of the joint posterior that depend on each parameter.

Full conditional pdf of  $\lambda$ :

$$\begin{aligned}
 p(\lambda | m, \phi, \mathbf{x}) & \propto \lambda^{x_{1,\star}(m)+\alpha-1} \exp[-\lambda (m + \beta)] \\
 & \propto \text{Gamma}(\alpha + x_{1,\star}(m), m + \beta).
 \end{aligned}$$

Full conditional pdf of  $\phi$ :

$$\begin{aligned} p(\phi | m, \lambda, \mathbf{x}) &\propto \phi^{x_{2,*}(m)+\gamma-1} \exp[-\phi(n-m+\delta)] \\ &\propto \text{Gamma}(\gamma + x_{2,*}(m), n - m + \delta). \end{aligned}$$

Full conditional pmf of  $m \in \{1, 2, \dots, n\}$ :

$$p(m | \lambda, \phi, \mathbf{x}) = \frac{1}{c} \cdot f(m, \lambda, \phi, \mathbf{x})$$

where

$$c = \sum_{k=1}^n f(k, \lambda, \phi, \mathbf{x})$$

and

$$\begin{aligned} f(m, \lambda, \phi, \mathbf{x}) &= \lambda^{x_{1,*}(m)+\alpha-1} \exp[-\lambda(m+\beta)] \\ &\quad \cdot \phi^{x_{2,*}(m)+\gamma-1} \exp[-\phi(n-m+\delta)] \\ &\quad \cdot i_{\{1,2,\dots,n\}}(m). \end{aligned}$$

We can use the inverse-cdf method to sample from  $p(m | \lambda, \phi, \mathbf{x})$ : just tabulate the pmf.

After the burn-in period (first  $K$  samples, say), we collect  $N$  draws  $(m^{(K)}, \lambda^{(K)}, \phi^{(K)})$ ,  $(m^{(K+1)}, \lambda^{(K+1)}, \phi^{(K+1)})$ ,  $\dots$ ,

$(m^{(K+N-1)}, \lambda^{(K+N-1)}, \phi^{(N+K-1)})$ . Then, for example, we can estimate the change point  $m$  simply as

$$\hat{m} = \frac{1}{N} \sum_{i=0}^{N-1} m^{(K+i)} \approx \text{the posterior mean of } p(m|\mathbf{x}).$$

# Can We Apply Grouping and Collapsing Ideas Here?

$$p(m, \phi | \mathbf{x}) = \frac{p(\phi, m, \lambda, \mathbf{x})}{p(\lambda | m, \phi, \mathbf{x})}$$

$\underbrace{\quad}_{\alpha}$   
 keep all the terms  
 containing  $\phi, m, \lambda$

$$\begin{aligned} & \frac{f(m, \lambda, \phi, \mathbf{x})}{\frac{(m+\beta)^{x_{1,*}(m)+\alpha}}{\Gamma(x_{1,*}(m)+\alpha)} \cdot \lambda^{x_{1,*}(m)+\alpha-1} \exp[-\lambda(m+\beta)]} \\ &= \frac{\lambda^{x_{1,*}(m)+\alpha-1} \exp[-\lambda(m+\beta)]}{\frac{(m+\beta)^{x_{1,*}(m)+\alpha}}{\Gamma(x_{1,*}(m)+\alpha)} \cdot \lambda^{x_{1,*}(m)+\alpha-1} \exp[-\lambda(m+\beta)]} \\ & \quad \cdot \phi^{x_{2,*}(m)+\gamma-1} \exp[-\phi \cdot (n-m+\delta)] \cdot i_{\{1,2,\dots,n\}}(m) \\ &= \frac{\Gamma(x_{1,*}(m)+\alpha)}{(m+\beta)^{x_{1,*}(m)+\alpha}} \cdot \phi^{x_{2,*}(m)+\gamma-1} \exp[-\phi \cdot (n-m+\delta)] \\ & \quad \cdot i_{\{1,2,\dots,n\}}(m) \stackrel{\Delta}{=} F(m, \phi, \mathbf{x}). \end{aligned}$$

Is  $p(\phi | m, \mathbf{x})$  a standard pdf? Yes, we recognize it in the table of distributions:

$$\begin{aligned} p(\phi | m, \mathbf{x}) & \propto \phi^{x_{2,*}(m)+\gamma-1} \exp[-\phi(n-m+\delta)] \\ & \propto \text{Gamma}(\gamma + x_{2,*}(m), n-m+\delta) \end{aligned}$$

which is the same as  $p(\phi | m, \lambda, \mathbf{x})$  — what can we conclude from that? The pmf  $p(m | \phi, \mathbf{x})$  is

$$p(m | \phi, \mathbf{x}) = \frac{F(m, \phi, \mathbf{x})}{\sum_{k=1}^n F(k, \phi, \mathbf{x})}.$$

So, we could just cycle between the following two steps:

- Draw a  $\phi^{(t+1)}$  from  $p(\phi | m^{(t)}, \mathbf{x})$ ,
- Draw an  $m^{(t+1)}$  from  $p(m | \phi^{(t+1)}, \mathbf{x})$ .

This is a *collapsed Gibbs sampler* (where  $\lambda$  has been integrated out). We could keep  $\lambda$  and integrate  $\phi$  out, using analogous arguments.

# Non-Standard Distributions

It may happen that one or more full conditionals is not a standard distribution. What do we do then? Try

- inverse cdf method, grid method, rejection sampling, composition etc.
- some approximation (e.g. a normal or  $t$ -distribution approximation around the MAP estimate, see handout # 4).
- more general MCMC algorithms: Metropolis, Metropolis-Hastings (M-H), or slice sampler.

# M-H Algorithm and Why it Might Work

We first consider the scalar case, for simplicity. (An extension to the vector case is trivial.)

- **Goal:** Simulate from a given distribution  $p(\theta)$  [ $\equiv p(\theta|\mathbf{x})$  in the Bayesian context].
- **Strategy:** Simulate an MC so that the limiting distribution is  $p(\theta)$ .
- **Algorithm:**

Start from a number  $\theta_0$  within the support of  $p(\theta)$ ;

**Step 1:** Draw a number  $\theta_*$  from the proposal distribution  $J(\theta_* | \theta^{(t)})$ ;

**Step 2:** Calculate the M-H ratio:

$$r = \frac{p(\theta_*)J(\theta^{(t)}|\theta_*)}{p(\theta^{(t)})J(\theta_*|\theta^{(t)})};$$

**Step 3:**

$$\theta^{(t+1)} = \begin{cases} \theta_*, & \text{with probability } p = \min\{1, r\} \\ \theta^{(t)}, & \text{with probability } 1 - p \end{cases} .$$

- **Remark 1:** Possible choices of the proposal distribution (*the convergence rate critically depends upon these choices*):



- for discrete  $p(\theta)$ , we may choose simple symmetric random walk:

$$J(\theta_\star | \theta^{(t)}) = \begin{cases} \theta^{(t)} + 1, & \text{with probability } p_{\text{ssrw}} = \frac{1}{2} \\ \theta^{(t)} - 1, & \text{with probability } 1 - p_{\text{ssrw}} = \frac{1}{2} \end{cases} .$$

- for continuous  $p(\theta)$ , choose a random-walk Gaussian proposal pdf:

$$J(\theta_\star | \theta^{(t)}) = \mathcal{N}\left(\theta | \theta^{(t)}, \underbrace{\sigma^2}_{\text{tuning parameter}}\right).$$

In both above examples, the proposal distributions are symmetric:

$$J(\theta_\star | \theta^{(t)}) = J(\theta^{(t)} | \theta_\star)$$

and, consequently, the M-H ratio simplifies to

$$r = \frac{p(\theta_\star)}{p(\theta^{(t)})}$$

which corresponds to the *Metropolis algorithm* (invented in the 1950s).

- **Remark 2:** How to do Step 3? Simulate a random number  $u \sim \text{uniform}(0, 1)$  and select

$$\theta^{(t+1)} = \begin{cases} \theta_\star, & \text{if } u < r \\ \theta^{(t)}, & \text{if } u \geq r \end{cases}$$

$\implies$  jump occurs with probability  $\min\{1, r\}$ .

- **Remark 3:** How do we know that the chain has converged?  
We can do the following:

$$\theta^{(0)}, \theta^{(1)}, \dots, \underbrace{\theta^{(l)}, \dots, \theta^{(j)}}_{\text{histogram}}, \underbrace{\theta^{(j+1)}, \dots, \theta^{(m)}}_{\text{histogram}}, \dots, \theta^{(n)}.$$

Or perhaps use autocorrelation function? These are easy to do in the one-dimensional case considered here.

- **Remark 4: Why M-H will converge to  $p(\theta)$ .**  
First, note that  $J(\theta_\star | \theta^{(t)})$  is the probability of proposed probability, so it is *different* from the actual transition probability  $\mathcal{P}(\theta_\star | \theta^{(t)})$  because an acceptance-rejection step is involved. We compute the transition probability as

$$\mathcal{P}(\theta_\star | \theta^{(t)}) = J(\theta_\star | \theta^{(t)}) \cdot \min \left\{ 1, \frac{p(\theta_\star) J(\theta^{(t)} | \theta_\star)}{p(\theta^{(t)}) J(\theta_\star | \theta^{(t)})} \right\}.$$

Therefore, if we start from  $p(\theta^{(t)})$ , we get

$$\begin{aligned}
 & p(\theta^{(t)}) \mathcal{P}(\theta_\star | \theta^{(t)}) \\
 &= p(\theta^{(t)}) J(\theta_\star | \theta^{(t)}) \min \left\{ 1, \frac{p(\theta_\star) J(\theta^{(t)} | \theta_\star)}{p(\theta^{(t)}) J(\theta_\star | \theta^{(t)})} \right\} \\
 &= \min \left\{ p(\theta^{(t)}) J(\theta_\star | \theta^{(t)}), p(\theta_\star) J(\theta^{(t)} | \theta_\star) \right\} \\
 &= p(\theta_\star) J(\theta^{(t)} | \theta_\star) \min \left\{ \frac{p(\theta^{(t)}) J(\theta_\star | \theta^{(t)})}{p(\theta_\star) J(\theta^{(t)} | \theta_\star)}, 1 \right\} \\
 &= p(\theta_\star) \mathcal{P}(\theta^{(t)} | \theta_\star).
 \end{aligned}$$

Therefore, in the discrete state-space case [with state space  $\{\theta_1, \theta_2, \dots, \theta_M\}$ ], we have

$$\underbrace{\sum_{i=1}^M p(\theta_i) \mathcal{P}(\theta_j | \theta_i)}_{\text{total probability}} = \sum_{i=1}^M p(\theta_j) \mathcal{P}(\theta_i | \theta_j) = p(\theta_j)$$

where the last equality follows from

$$\sum_{i=1}^M \mathcal{P}(\theta_i | \theta_j) = 1.$$

Putting the above equations in the matrix form (for  $j =$

$1, 2, \dots, J$ ), we obtain

$$\mathbf{p}^T \mathcal{P} = \mathbf{p} \quad (\text{equation of full balance})$$

i.e.  $\mathbf{p} = [p(\theta_1), p(\theta_2), \dots, p(\theta_M)]^T$  is the stationary distribution!

- **Remark 4'**: The above proof implies that the following condition must hold:

$$J(\theta_\star | \theta) > 0 \quad \text{if and only if} \quad J(\theta | \theta_\star) > 0.$$

Otherwise, the sampler will not converge to the desired stationary distribution. This is the only serious restriction on the proposal distribution.

- **Remark 5**: The pdf/pmf  $p(\theta)$  needs to be known only up to a proportionality constant, since this constant cancels out when computing the M-H ratio.

# M-H Algorithm: A Simple Example

- Target pdf:  $p(\theta) = \mathcal{N}(\mu, \sigma^2)$ ;
- Proposal pdf:  $J(\theta_\star | \theta^{(t)}) = \mathcal{N}(\theta^{(t)}, \tau^2)$ ;

Therefore, since  $J(\theta_\star | \theta^{(t)}) = J(\theta^{(t)} | \theta_\star)$  (symmetry, i.e. the Metropolis case), we have

$$r = \frac{p(\theta_\star)}{p(\theta^{(t)})} = \exp \left[ -\frac{1}{2\sigma^2} (\theta_\star - \mu)^2 + \frac{1}{2\sigma^2} (\theta^{(t)} - \mu)^2 \right].$$

**Important side comment:** All the above expressions are nice and fine for writing, but when implementing, in general, do all your computations in the log scale.

# Random-walk Metropolis

- Here, we generate candidates  $\theta_*$  using random walk. (Note that we trivially switch to the vector notation.)
- **A popular choice for proposal pdf.** Pick the proposal distribution to be Gaussian centered at the current draw:

$$J(\theta_*|\theta^{(t)}) = \mathcal{N}(\theta^{(t)}, V).$$

- Comments:
  - The above  $J(\cdot|\cdot)$  is symmetric, hence this is a Metropolis sampler.
  - It may be difficult to choose a good  $V$ :
    - If  $V$  too small, it takes a long time to explore the parameter space.
    - If  $V$  too large, jumps to extremes are less likely to be accepted; consequently, the chain stays in the same place too long.
    - The ideal  $V$ : posterior variance. Bad choice of  $V$  leads to poor mixing (we have introduced a notion of poor mixing before).

- **Of course, if the problem is difficult, we may not be able to find any  $V$  that works (i.e. explores the entire parameter space within our lifetimes)!**
- Good acceptance rate (if we can get it): 20% to 50%, in particular,
  - $\min\{1, r\} \in [40\%, 50\%]$  in the scalar case;
  - $\min\{1, r\} \in [20\%, 30\%]$  in the vector case.

but even following this suggestion does not guarantee good performance, since performance will also depend on difficulty (i.e. “nastiness” of the stationary distribution).

# Independence Sampler

- Here, the proposal distribution  $J(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})$  does not depend on  $\boldsymbol{\theta}^{(t)}$ .
- Just choose a distribution  $\tilde{p}(\boldsymbol{\theta})$  and draw samples from it.
- Here,  $\tilde{p}(\boldsymbol{\theta}_*) = J(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)}) \neq J(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}_*) = \tilde{p}(\boldsymbol{\theta}^{(t)})$  (it is good to see some M-H example that is not a Metropolis sampler) and, therefore, we have

$$r = \frac{p(\boldsymbol{\theta}_*) \tilde{p}(\boldsymbol{\theta}^{(t)})}{p(\boldsymbol{\theta}^{(t)}) \tilde{p}(\boldsymbol{\theta}_*)} = \frac{\phi(\boldsymbol{\theta}_*)}{\phi(\boldsymbol{\theta}^{(t)})}$$

where

$$\phi(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta})}$$

is simply the importance ratio.

- This sampler is an alternative to rejection sampling and importance sampling.
- Its performance depends on how well  $\tilde{p}(\boldsymbol{\theta})$  approximates  $p(\boldsymbol{\theta})$ .
- This sampler has a “global” nature — in contrast to the random-walk Metropolis algorithm, which tends to do more “local” explorations.



- Typically [e.g. if the support of  $p(\boldsymbol{\theta})$  is infinite],  $\tilde{p}(\boldsymbol{\theta})$  should be chosen to have heavy tails. In this case, we can use a (multivariate)  $t$  distribution for  $\tilde{p}(\boldsymbol{\theta})$ ; the smaller the degree-of-freedom parameter of the  $t$  distribution, the heavier the tails.

A useful reference:

J.S. Liu, "Metropolized independent sampling with comparisons to rejection sampling and importance sampling," *Statistics and Computing*, vol. 6, pp. 113–119, Jun. 1996.

# (Univariate) Slice Sampler

Consider now sampling a random variable  $\phi$  from a nonstandard  $p(\phi) \propto h(\phi)$ .

## (Seemingly Counter-Intuitive!) Idea:

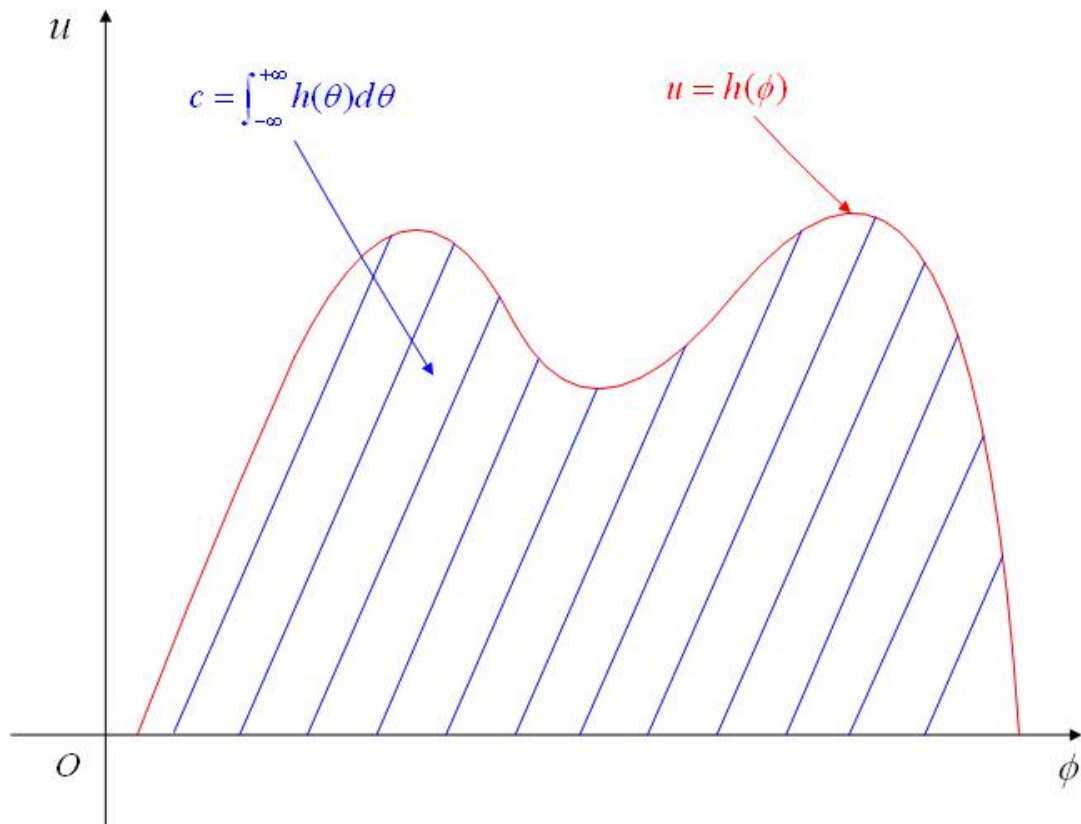
- Invent a convenient bivariate distribution for, say,  $\phi$  and  $u$ , with marginal pdf for  $\phi$  specified by  $h(\phi)$ .
- Then, use Gibbs sampling to make

$$(\phi^{(0)}, u^{(0)}), (\phi^{(1)}, u^{(1)}), (\phi^{(2)}, u^{(2)}), \dots, (\phi^{(T)}, u^{(T)}).$$

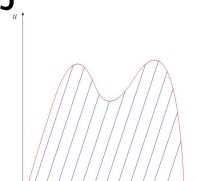


Create an auxiliary variable  $u$  just for convenience!

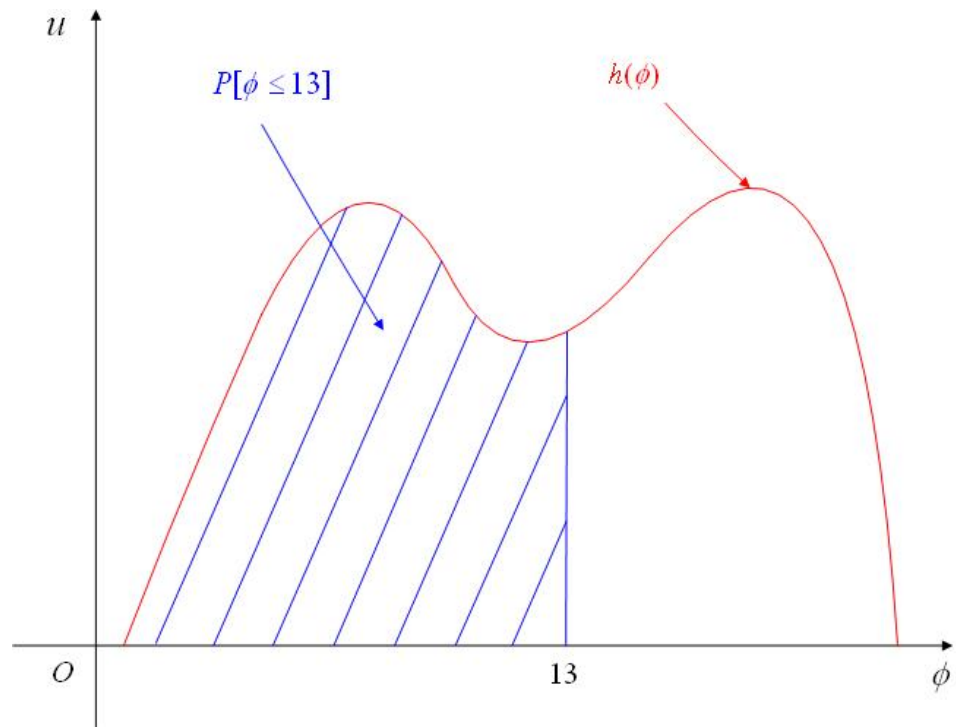
# (Univariate) Slice Sampler



“Invent” a joint distribution for  $\phi$  and  $u$  by declaring it to be

uniform on  :

$$p(\phi, u) = \begin{cases} \frac{1}{c}, & 0 < u < h(\phi) \\ 0, & \text{otherwise} \end{cases} \propto i_{(0, h(\phi))}(u).$$



With this joint pdf,  $P[\phi \leq 13] = \int_{-\infty}^{13} \frac{h(\phi)}{c} d\phi$ .



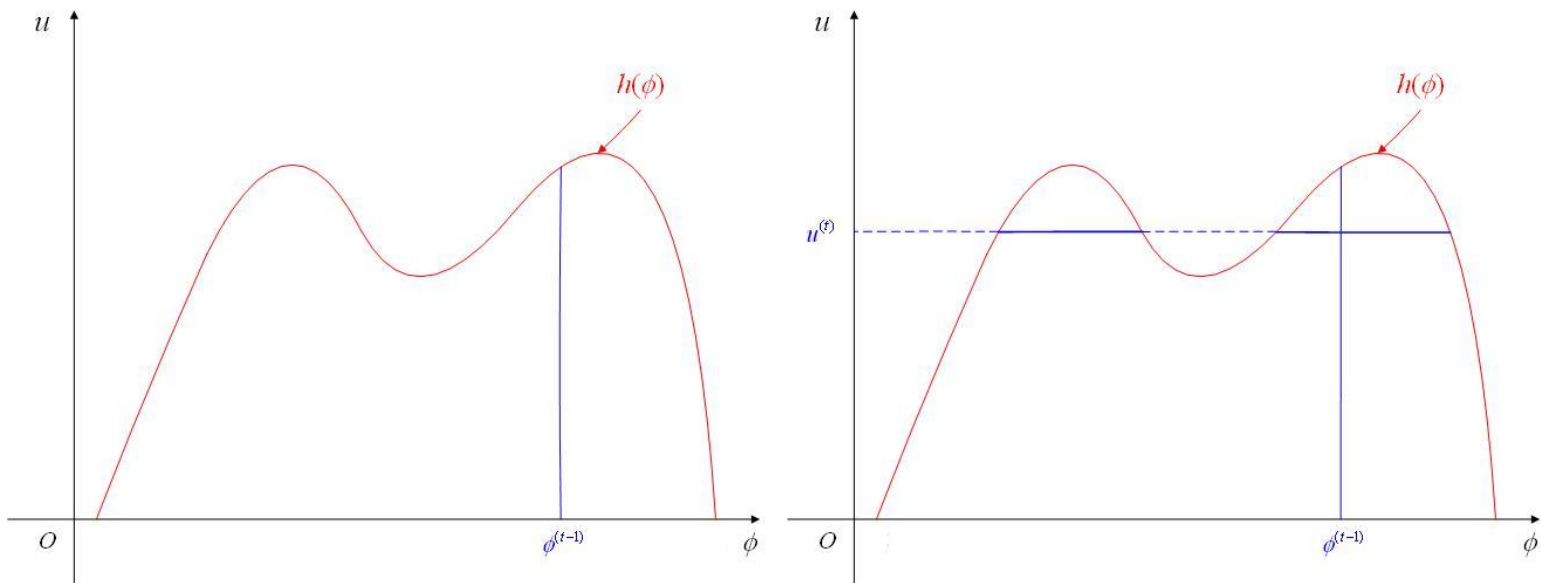
The marginal pdf of  $\phi$  is indeed specified by  $h(\phi) \implies$  if we figure out how to do Gibbs sampling, we know how to generate a  $\phi$  from  $h(\phi)$ .

# Gibbs Sampler is Easy in This Case!

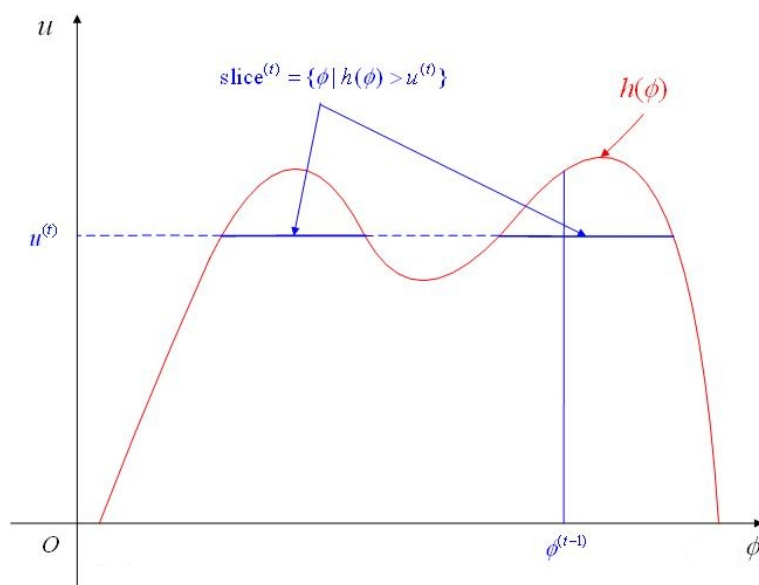
$$p(u | \phi) = \text{uniform}(0, h(\phi))$$

$$p(\phi | u) = \text{uniform on } \underbrace{\{\phi | h(\phi) > u\}}_{\text{"slice"}}$$

**Step 1: Given  $\phi^{(t-1)}$ , sample  $u^{(t)} \sim \text{uniform}(0, h(\phi^{(t-1)}))$**



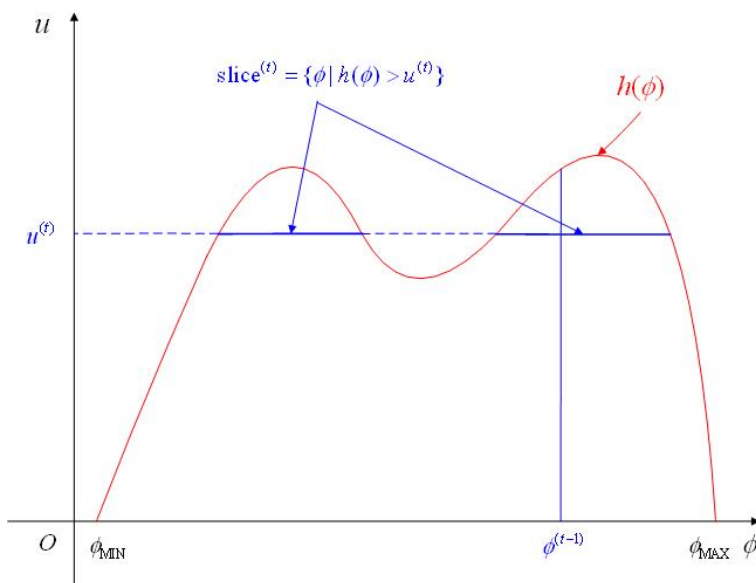
**Step 2: Given  $u^{(t)}$ , sample  $\phi^{(t)}$  Uniform from slice<sup>(t)</sup>**



If we can algebraically solve  $h(\phi) = u^{(t)}$ , our task is easy. What if not?

## Step 2 implementation using the rejection method

When we have band bounds on  $\phi$ , say  $\phi_{\text{MIN}} \leq \phi \leq \phi_{\text{MAX}}$



generate i.i.d. values  $\phi$  from  $\text{uniform}(\phi_{\text{MIN}}, \phi_{\text{MAX}})$  until we produce a  $\phi$  in the slice [i.e.  $h(\phi) > u^{(t)}$ ], which we then accept as  $\phi^{(t)}$ . This is nothing but rejection sampling!

**Note:** For multivariate extensions of the slice sampler (particularly the “shrinkage idea”), see

R.M. Neal, “Slice sampling,” *Ann. Statist.*, vol. 31, pp. 705–741, June 2003.

# MCMC: Final General Comment

MCMC will in practice (i.e. when dealing with real problems rather than toy examples) yield only a “best guess” of the real posterior. This is all that we can expect.

# A Bit About Long-run Behavior of MCMC Samplers

Suppose that we have an “ergodic” Markov chain  $\{\theta^{(t)} \mid t = 0, 1, 2, \dots\}$  generated by the transition kernel  $\mathcal{P}$  with stationary distribution  $p(\theta)$ .

Suppose that we wish to evaluate

$$G = \int_{\Omega} g(\theta) p(\theta) d\theta < \infty$$

which we discussed already before. Recall our basic estimate of  $G$ :

$$\hat{G}_N = \frac{1}{N} \cdot \underbrace{\sum_{t=1}^N g(\theta^{(t)})}_{S_N(g)}.$$

As seen in Theorem 6, under certain regularity conditions, we have

$$\frac{1}{N} S_N(g) \rightarrow G$$

or, in words, *time averages converge to the ensemble average*.

The asymptotic variance of the above estimate:

$$\text{var} \left[ \frac{1}{N} S_N(g) \right]$$



where  $\text{var}$  is taken with respect to the transition kernel  $p_{\cdot}$ , assuming stationarity, i.e. that  $\theta^{(1)}$  comes from  $p(\theta^{(1)})$ . Define

$$v(g, \mathcal{P}) \triangleq \lim_{n \rightarrow \infty} N \cdot \text{var} \left[ \frac{1}{N} S_N(g) \right].$$

Under certain regularity conditions,

$$\sqrt{N} \left( \frac{1}{N} S_N(g) - G \right) \xrightarrow{d} \mathcal{N}(0, v(g))$$

implying that  $v(g, \mathcal{P})$  determines the asymptotic accuracy of  $\frac{1}{N} S_N(g)$ . Note that

$$\begin{aligned} N \cdot \text{var} \left[ \frac{1}{N} S_N(g) \right] &= \frac{1}{N} \cdot \left\{ \underbrace{\text{var}[g(\theta^{(1)})]}_{\sigma^2} + \underbrace{\text{var}[g(\theta^{(2)})]}_{\sigma^2} + \dots \right. \\ &\quad \left. + \underbrace{\text{var}[g(\theta^{(N)})]}_{\sigma^2} + \sum_{l=1}^N \sum_{\substack{m=1 \\ m \neq l}}^N \underbrace{\text{cov}[g(\theta^{(l)}), g(\theta^{(m)})]}_{\rho_{l-m} \cdot \sigma^2} \right\} \\ &= \sigma^2 \left[ 1 + 2 \sum_{j=1}^{n-1} \left( 1 - \frac{j}{n} \right) \rho_j \right] \end{aligned}$$

where

$$\begin{aligned}\sigma^2 &= \text{var}[g(\theta^{(t)})] \\ \rho_{l-m} &= \rho_{m-l} = \frac{\text{cov}[g(\theta^{(l)}), g(\theta^{(m)})]}{\sigma^2} \\ &= \text{correlation coeff. between } g(\theta^{(l)}) \text{ and } g(\theta^{(m)}) .\end{aligned}$$

Suppose that we wish to compare two ergodic Markov chains with the same stationary distribution  $p(\theta)$ . Then, the one for which the  $\rho_j$ s are smaller will have more accurate estimates (because  $\sigma^2$  will be the same for both chains).

To summarize, we look for an MCMC sampler with as small autocorrelation among its draws as possible.

**Note:** As  $N \rightarrow \infty$ ,

$$N \cdot \text{var} \left[ \frac{1}{N} S_N(g) \right] \approx \sigma^2 \left( 1 + 2 \sum_{j=1}^{\infty} \rho_j \right)$$

and, therefore,

$$v(g, p_{\cdot, \cdot}) = \sigma^2 \left( 1 + 2 \sum_{j=1}^{\infty} \rho_j \right)$$

$\implies N \cdot \text{var} \left[ \frac{1}{N} S_N(g) \right] \longrightarrow v(g, \mathcal{P})$  (instead of just  $\sigma^2$ , which would have been the case if the  $\theta^{(t)}$ s were all independent).

Here, the factor  $1 + 2 \sum_{j=1}^{\infty} \rho_j$  is the penalty that we pay for using dependent samples. Define *integrated autocorrelation time* of  $g$  as

$$\tau_{\text{int}}(g, \mathcal{P}) = \frac{1}{2} + \sum_{j=1}^{\infty} \rho_j$$

a definition taken from the physics literature. Now, we have

$$\text{var} \left[ \frac{1}{N} S_N(g) \right] = \sigma^2 / \left( \frac{N}{2\tau_{\text{int}}(g, \mathcal{P})} \right)$$

which is approximately the variance that we would have achieved if we had  $\frac{N}{2\tau_{\text{int}}(g, \mathcal{P})}$  independent observations. Therefore, the quantity

$$\frac{N}{2\tau_{\text{int}}(g, \mathcal{P})}$$

is called the *effective sample size*. We can use (estimated or, in some cases, analytical) integrated autocorrelation time or effective sample size to assess the performance of a given sampler.

# A FEW BITS AND PIECES ON MCMC

## Combining M-H Samplers

Suppose that we have a bunch of M-H proposals  $J_j(\cdot | \cdot)$ ,  $j = 1, 2, \dots, J$  for sampling from the same density  $p(\boldsymbol{\theta})$  and let us assign a probability  $q_j$  to each proposal. Define  $\mathbf{q} = [q_1, q_2, \dots, q_J]^T$ . Let us also denote by  $\mathcal{P}_j(\cdot | \cdot)$  the M-H transition kernel corresponding to  $J_j(\cdot | \cdot)$ ,  $j = 1, 2, \dots, J$ .

Consider the following transition kernel:

$$[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]^{\text{MT}} = \sum_{j=1}^J q_j \mathcal{P}_j(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})$$

where MT stands for mixture of **M-H Transition** kernels. How do we construct such a sampler?

Use composition sampling: pick the transition kernel  $\mathcal{P}_j(\cdot | \cdot)$  with probability  $q_j$ ,  $j = 1, 2, \dots, J$ . In particular, here are the details of the MT scheme:

Start from a value  $\boldsymbol{\theta}_0$  within the support of  $p(\boldsymbol{\theta})$ ;

**Step 1:** Draw a  $j \in \{1, 2, \dots, J\}$  with probability  $q_j$ ;

**Step 2:** Draw a  $\boldsymbol{\theta}_*$  from the  $j$ th proposal distribution  $J_j(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})$ ;

**Step 3:** Calculate the M-H ratio:

$$r_j^{\text{MT}} = \frac{p(\boldsymbol{\theta}_*) J_j(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}_*)}{p(\boldsymbol{\theta}^{(t)}) J_j(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})};$$

**Step 4:**

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}_*, & \text{with probability } p = \min\{1, r_j^{\text{MT}}\} \\ \boldsymbol{\theta}^{(t)}, & \text{with probability } 1 - p \end{cases} .$$

## Combining M-H Samplers (cont.)

**An alternative way of combining M-H samplers.** Construct a mixture of  $J$  proposals:

$$[J(\cdot | \cdot)]^{\text{MP}} = \sum_{j=1}^J q_j J_j(\cdot | \cdot)$$

where MP stands for mixture of **M-H P**roposals. Here, we use  $[J(\cdot | \cdot)]^{\text{MP}}$  at each iteration.

Start from a value  $\boldsymbol{\theta}_0$  within the support of  $p(\boldsymbol{\theta})$ ;

**Step 1:** Draw a  $j \in \{1, 2, \dots, J\}$  with probability  $q_j$ ;

**Step 2:** Draw a  $\boldsymbol{\theta}_*$  from the proposal distribution  $J_j(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})$ ;

**Step 3:** Calculate the M-H ratio:

$$r^{\text{MP}} = \frac{p(\boldsymbol{\theta}_*) \sum_{j=1}^J q_j J_j(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}_*)}{p(\boldsymbol{\theta}^{(t)}) \sum_{j=1}^J q_j J_j(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})};$$

**Step 4:**

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}_*, & \text{with probability } p = \min\{1, r^{\text{MP}}\} \\ \boldsymbol{\theta}^{(t)}, & \text{with probability } 1 - p \end{cases} .$$

What to use in practice: mixture of M-H updates or mixture of M-H proposals? It can be shown that it is better to use the mixture of M-H proposals with transition kernel  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]_{\text{MP}}$  than  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]_{\text{MT}}$ , but

- using  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]_{\text{MP}}$  requires evaluation of all the proposal densities at every iteration, even though we sample from one density only;
- $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]_{\text{MT}}$  is not as expensive as  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]_{\text{MP}}$ .

**An idea for showing that  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]_{\text{MP}}$  is better:** Show that  $r^{\text{MP}} > r^{\text{MT}} \implies$  draws from the MP sampler will be less correlated than those from the MT sampler!



Hence,

- In some cases, it may be more beneficial to
  - run the cheaper chain with the transition kernel  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]^{\text{MT}}$  longercompared with
  - a shorter run *taking the same amount of time* of the more “efficient” chain with the transition kernel  $[\mathcal{P}(\boldsymbol{\theta}_* | \boldsymbol{\theta}^{(t)})]^{\text{MP}}$ .

Recall the definition of the *effective sample size*:

$$\frac{N}{2 \tau_{\text{int}}(g, \mathcal{P})}$$

Equivalently, we may have

$$\frac{N^{\text{MP}}}{2 \tau_{\text{int}}^{\text{MP}}(g, \mathcal{P})} < \frac{N^{\text{MT}}}{2 \tau_{\text{int}}^{\text{MT}}(g, \mathcal{P})}$$

even though  $\tau_{\text{int}}^{\text{MP}}(g, \mathcal{P}) < \tau_{\text{int}}^{\text{MT}}(g, \mathcal{P})$ .

# A Bit About General Conditional Sampling (Ch. 8 in Liu)

Gibbs is the first and simplest conditional-sampling method. But, Gibbs is restricted by the parameterization/coordinate system.

Formulate “a move” in the space as a point being transformed/mapped to another point. Possible moves constitute a set of transformations.

A Gibbs-sampler move for fixed  $\theta_2$ : draw a  $\theta_1^{(t+1)}$  from

$$p_{\theta_1 | \theta_2}(\theta_1 | \theta_2^{(t)}) \propto p_{\theta_1, \theta_2}(\theta_1, \theta_2^{(t)}).$$

which corresponds to a move:

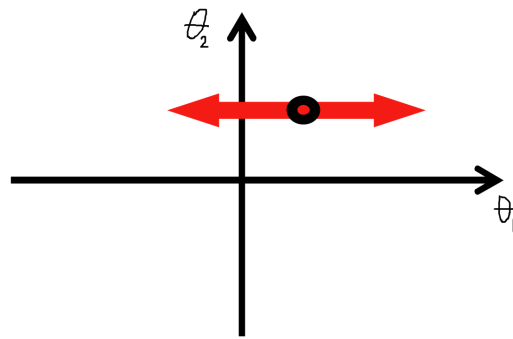
$$\theta_1^{(t)} \rightarrow \theta_1^{(t+1)}$$

where  $\theta_1^{(t)}$  is the value of  $\theta_1$  from the previous cycle. This can be seen as

$$\theta_1^{(t)} \rightarrow \theta_1^{(t)} + c$$

where  $c$  is drawn from the following pdf/pmf:

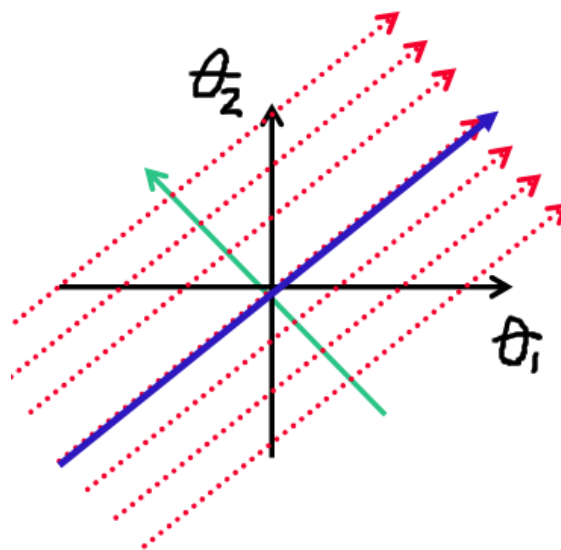
$$p_c\left(c | (\theta_1^{(t)}, \theta_2^{(t)})\right) \propto p_{\theta_1, \theta_2}(\theta_1^{(t)} + c, \theta_2^{(t)}).$$



This move is fairly limited — Gibbs jumps only in directions parallel to the coordinate axes.

How about some other direction (corresponding to rotating the coordinate system, say), such as

$$(\theta_1, \theta_2) \rightarrow (\theta_1 + c, \theta_2 + c)?$$



This is effectively a reparametrization and we need to make sure that  $p(\boldsymbol{\theta})$  is invariant under this new move. To do that,

we draw  $c$  from

$$p_c(c | (\theta_1, \theta_2)) \propto p_{\theta_1, \theta_2}(\theta_1 + c, \theta_2 + c).$$

## Let's be Even More Creative ...

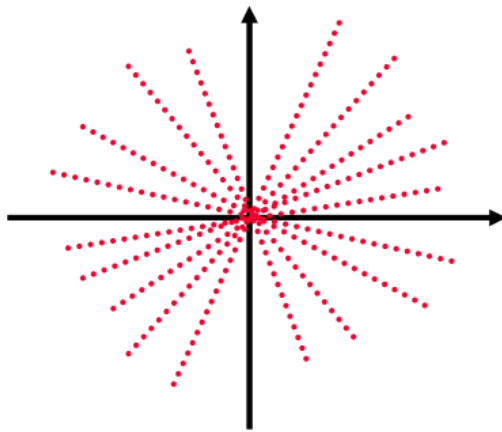
Try

$$(\theta_1, \theta_2) \rightarrow \gamma \cdot (\theta_1, \theta_2) \quad (1)$$

or, perhaps,

$$(\theta_1, \theta_2) \rightarrow (\theta_1, \theta_2) A$$

where  $A$  is an orthonormal matrix. What are the distributions of  $\gamma$  and  $A$  so that  $p_{\theta}(\boldsymbol{\theta})$  is invariant under these moves?



In general, we have a *group of transformations*  $\Gamma$  to represent possible moves. Hence, pick

$$\gamma \in \Gamma$$

and apply a move  $\boldsymbol{\theta} \rightarrow \gamma(\boldsymbol{\theta})$ . In the scaling case (1), we draw  $\gamma$  from a  $p_{\gamma|\boldsymbol{\theta}}(\gamma|\boldsymbol{\theta})$  so that

$$\boldsymbol{\theta}' = \gamma \boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}(\cdot) \quad \text{if } \boldsymbol{\theta} \sim p_{\boldsymbol{\theta}}(\cdot).$$

**Note:** the group of transformations partitions the space into “orbits.”

## A Theorem

**Theorem 7.** Suppose that  $\Gamma = \{\text{all } \gamma\}$  forms a *locally compact group* and let  $H(d\gamma)$  be its unimodular *Haar measure*. If  $\theta \sim p_\theta(\theta)$  and

$$\gamma \sim p_{\gamma|\theta}(\gamma | \theta) \propto p_\theta(\gamma(\theta)) \cdot \left| \frac{\partial \gamma(\theta)}{\partial \theta^T} \right| \cdot H(d\gamma) \quad (2)$$

then  $\gamma(\mathbf{x})$  follows a distribution  $p_\theta(\cdot)$ .

The distribution (2) is “independent” of the position  $\theta$ .

**Note:** A left-invariant Haar measure satisfies:

$$H(\underbrace{\gamma B}_{\substack{\gamma \text{ acting on} \\ \text{every element of } B}}) = H(B), \quad \forall \gamma, B.$$

**Example:**

$$p_z(\mathbf{z}) \propto e^{-\beta \cdot H(\mathbf{z})}.$$

Update

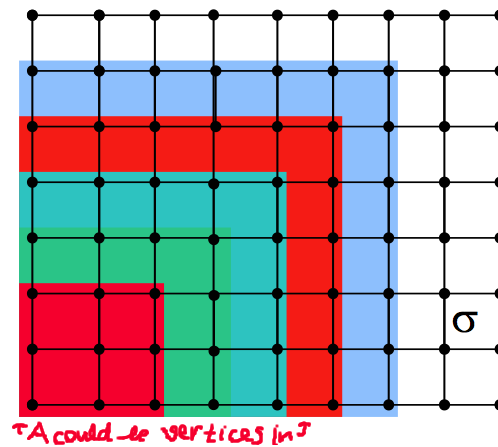
$$\mathbf{z} \rightarrow \mathbf{z} + t \mathbf{i}_A(\mathbf{z})$$

where  $\mathbf{z} \equiv$  all lattice points (pixels)  $j$ . Here,  $t$  should be drawn from

$$p_t(t) \propto e^{-\beta \cdot H(\mathbf{z} + t \mathbf{i}_A(\mathbf{z}))}$$

where  $A$  is a subset of the pixel set and  $\mathbf{i}_A(\mathbf{z})$  is the vector of indicator functions, i.e. the  $j$ th element of  $\mathbf{i}_A(\mathbf{z})$  is one if  $j \in A$  and zero otherwise.

Requirement: if  $\mathbf{z} \sim p_{\mathbf{z}}(\cdot)$ , then  $\mathbf{z} + t \mathbf{i}_A(\mathbf{z})$  also follows  $p_{\mathbf{z}}(\cdot)$ .



## Example: General Conditional Sampling

Suppose that the observations  $\mathbf{y}_t$  are i.i.d. from a multivariate  $t$  distribution (see the table of distributions):

$$p(\underbrace{\mathbf{y}_i}_{p \times 1} \mid \underbrace{\boldsymbol{\mu}}_{p \times 1}, \underbrace{\boldsymbol{\Sigma}}_{p \times p}) = t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, 2, \dots, N$$

where  $\nu \equiv$  (known) degrees of freedom. The table of distributions yields:

$$t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \text{const} \cdot |\boldsymbol{\Sigma}|^{-1/2} \cdot \left[ 1 + \frac{1}{\nu} \cdot (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right]^{-(\nu+p)/2}$$

which does not look friendly at all! Yet, this model can handle outliers well, because the multivariate  $t$  distribution can have heavy tails. So, it would be nice if we could find ML estimates of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  for the above model. These ML estimates cannot be found in a closed form.

**Fact:** if a  $p \times 1$  vector  $\mathbf{y}_i$  follows a  $t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , then we can



introduce missing data  $\mathbf{u}$  so that

$$\begin{aligned}
 p(\mathbf{y}_i | u_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u_i) \\
 &= \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}/u_i|}} \cdot \exp \left[ -\frac{1}{2} u_i \cdot \underbrace{d(\mathbf{y}_i, \boldsymbol{\mu}; \boldsymbol{\Sigma})}_{(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})} \right] \\
 &= \frac{1}{\sqrt{|2\pi \boldsymbol{\Sigma}|}} \cdot u_i^{p/2} \cdot \exp \left[ -\frac{1}{2} u_i \cdot d(\mathbf{y}_i, \boldsymbol{\mu}; \boldsymbol{\Sigma}) \right] \\
 p(u_i) &= \text{Gamma}(\nu/2, \nu/2) = \chi_\nu^2 / \nu \\
 &= \underbrace{\text{const}}_{\text{indep. of } u_i} \cdot u_i^{\nu/2-1} \cdot \exp(-\nu/2 \cdot u_i).
 \end{aligned}$$

Assume that we wish to estimate  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  and let us assign the standard Jeffreys' noninformative prior for these parameters:

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{p+1}{2}}.$$

Define

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}.$$

Now

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{u} \mid \mathbf{y}) \propto \left( \prod_{i=1}^N u_i^{(p+\nu)/2-1} \right) \cdot \exp \left( -\nu/2 \cdot \sum_{i=1}^N u_i \right) \\ \cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^N u_i \cdot d(\mathbf{y}_i, \boldsymbol{\mu}; \boldsymbol{\Sigma}) \right] \cdot |\boldsymbol{\Sigma}|^{-\frac{N+p+1}{2}}.$$

We can use Gibbs to sample from the above distribution. For this purpose, we need the following full conditionals:

- (a) Conditional on  $\mathbf{u}$  and  $\boldsymbol{\mu}$ , the posterior pdf of  $\boldsymbol{\Sigma}$  is an inverse-Wishart pdf:

$$p(\boldsymbol{\Sigma} \mid \mathbf{u}, \boldsymbol{\mu}, \mathbf{y}) \propto |\boldsymbol{\Sigma}|^{-\frac{N+p+1}{2}} \\ \cdot \exp \left\{ -\frac{1}{2} \operatorname{tr} \left[ \sum_{i=1}^N u_i (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \right] \right\} \\ = \text{Inv-Wishart}_N \left( \left\{ \sum_{i=1}^N u_i (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T \right\}^{-1} \right)$$

see your distribution table.

- Conditional on  $\mathbf{u}$  and  $\boldsymbol{\Sigma}$ , the posterior pdf of  $\boldsymbol{\mu}$  is a Gaussian

pdf:

$$p(\boldsymbol{\mu} | \mathbf{u}, \Sigma, \mathbf{y}) = \mathcal{N}\left(\underbrace{\frac{\sum_{i=1}^N u_i \mathbf{y}_i}{\sum_{i=1}^N u_i}}_{\triangleq \hat{\mathbf{y}}(\mathbf{u})}, \frac{\Sigma}{\sum_{i=1}^N u_i}\right)$$

- Conditional on  $\boldsymbol{\mu}$  and  $\Sigma$  (as well as  $\mathbf{y}$ , of course),  $u_i$  are mutually independent following

$$\begin{aligned} p(u_i | \boldsymbol{\mu}, \Sigma, \mathbf{y}) &\propto u_i^{(p+\nu)/2-1} \cdot \exp\left[-\frac{\nu + d(\mathbf{y}_i, \boldsymbol{\mu}; \Sigma)}{2} \cdot u_i\right] \\ &= \text{Gamma}\left(\frac{p + \nu}{2}, \frac{\nu + d(\mathbf{y}_i, \boldsymbol{\mu}; \Sigma)}{2}\right) \end{aligned}$$

see your distribution table.

## Gibbs Sampler for Inference on Mean Vector and Covariance Matrix of a Multivariate $t$ Distribution

Can we apply grouping and collapsing ideas here? Yes, we can

marginalize  $\boldsymbol{\mu}$  as follows:

$$p(\boldsymbol{\Sigma}, \mathbf{u} \mid \mathbf{y}) = \frac{p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{u} \mid \mathbf{y})}{p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mathbf{u}, \mathbf{y})}$$

$\propto$   
keep all the terms containing  $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{u}$

$$\left( \prod_{i=1}^N u_i^{(p+\nu)/2-1} \right) \cdot \exp \left( -\nu/2 \cdot \sum_{i=1}^N u_i \right)$$

$$\cdot \exp \left[ -\frac{1}{2} \sum_{i=1}^N u_i \cdot (\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right]$$

$$\cdot |\boldsymbol{\Sigma}|^{-\frac{N+p+1}{2}} \cdot |\boldsymbol{\Sigma} / \left( \sum_{i=1}^N u_i \right)|^{1/2}$$

$$\cdot \exp \left\{ \frac{1}{2} \cdot \left( \sum_{j=1}^N u_j \right) \cdot [\boldsymbol{\mu} - \hat{\mathbf{y}}(\mathbf{u})]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{\mu} - \hat{\mathbf{y}}(\mathbf{u})] \right\}$$

$=$   
choose  $\boldsymbol{\mu} = \hat{\mathbf{y}}(\mathbf{u})$

$$\left( \prod_{i=1}^N u_i^{(p+\nu)/2-1} \right) \cdot \exp \left( -\nu/2 \cdot \sum_{i=1}^N u_i \right)$$

$$\cdot \exp \left\{ -\frac{1}{2} \sum_{i=1}^N u_i \cdot [\mathbf{y}_i - \hat{\mathbf{y}}(\mathbf{u})]^T \boldsymbol{\Sigma}^{-1} [\mathbf{y}_i - \hat{\mathbf{y}}(\mathbf{u})] \right\}$$

$$\cdot |\boldsymbol{\Sigma}|^{-\frac{N+p+1}{2}} \cdot |\boldsymbol{\Sigma} / \left( \sum_{i=1}^N u_i \right)|^{1/2}$$

implying that

$$p(\Sigma | \mathbf{u}, \mathbf{y}) = \text{Inv-Wishart}_{N-1} \left( \left\{ \sum_{i=1}^N u_i [\mathbf{y}_i - \hat{\mathbf{y}}(\mathbf{u})] [\mathbf{y}_i - \hat{\mathbf{y}}(\mathbf{u})]^T \right\}^{-1} \right).$$

Therefore, we can sample a  $\Sigma$  from  $p(\Sigma | \mathbf{u}, \mathbf{y})$  (with  $\mu$  integrated out) rather than from the full conditional pdf in (a) and (slightly collapsed) Gibbs sampler consists of cycling between the following steps:

- Draw a  $\Sigma^{(t+1)}$  from  $p(\Sigma | \mathbf{u}^{(t)}, \mathbf{y})$

$$= \text{Inv-Wishart}_{N-1} \left( \left\{ \sum_{i=1}^N u_i^{(t)} [\mathbf{y}_i - \hat{\mathbf{y}}(\mathbf{u}^{(t)})] [\mathbf{y}_i - \hat{\mathbf{y}}(\mathbf{u}^{(t)})]^T \right\}^{-1} \right);$$

- Draw a  $\mu^{(t+1)}$  from  $p(\mu | \Sigma^{(t+1)}, \mathbf{u}^{(t)}, \mathbf{y})$   
 $= \mathcal{N} \left( \hat{\mathbf{y}}(\mathbf{u}^{(t)}), \Sigma^{(t+1)} / \left( \sum_{i=1}^N u_i^{(t)} \right) \right).$

**Together, the above two sampling steps yield a “grouped sample”  $(\Sigma^{(t+1)}, \mu^{(t+1)})$  from  $p(\mu, \Sigma | \mathbf{u}^{(t)}, \mathbf{y})$ .**

- Draw  $u_i^{(t+1)}$ ,  $i = 1, 2, \dots, N$  i.i.d. from  $p(u_i | \Sigma^{(t+1)}, \mu^{(t+1)}, \mathbf{y})$   
 $= \text{Gamma} \left( \frac{p+\nu}{2}, \frac{\nu + d(\mathbf{y}_i, \mu^{(t+1)}; \Sigma^{(t+1)})}{2} \right)$ , making  $\mathbf{u}^{(t+1)} = [u_1^{(t+1)}, u_2^{(t+1)}, \dots, u_N^{(t+1)}]^T$ .

**Note:** The samples  $\Sigma^{(t)}$  and  $\mathbf{u}^{(t)}$  are tightly coupled — if the starting values of  $u_i^{(t)}$  are large, the resulting sample of

$\Sigma$  tends to be large and vice versa. Recall that **P**arameter **eX**pansion helped solving this problem when we applied it to the EM iteration. Can we make a PX conditional-sampling move that will make an analogous improvement to the above Gibbs sampler?

Let us try this move:

$$(\Sigma, \mathbf{u}) \rightarrow (\Sigma/\alpha, \mathbf{u}/\alpha)$$

where, according to the above theorem:

$$\begin{aligned} p(\alpha | \Sigma, \boldsymbol{\mu}, \mathbf{u}, \mathbf{y}) &\propto p(\boldsymbol{\mu}, \Sigma/\alpha, \mathbf{u}/\alpha | \mathbf{y}) \cdot \alpha^{-\frac{p(p+1)}{2}-N} \cdot \alpha^{-1} \\ &\propto \left(\frac{1}{\alpha}\right)^{N(p+\nu)/2-N} \cdot \exp\left(-\frac{\nu}{2\alpha} \cdot \sum_{i=1}^N u_i\right) \\ &\quad \cdot \alpha^{\frac{N+p+1}{2}p} \cdot \alpha^{-\frac{p(p+1)}{2}-N} \cdot \alpha^{-1} \\ &= \alpha^{-\frac{N\nu}{2}-1} \cdot \exp\left(-\frac{\nu}{2\alpha} \cdot \sum_{i=1}^N u_i\right) \\ &= \text{Inv-}\chi^2\left(N\nu, \frac{1}{N} \sum_{i=1}^N u_i\right) \\ &= \frac{\nu \cdot \sum_{i=1}^N u_i}{X} \end{aligned}$$

where  $X$  is a  $\chi_{N\nu}^2$  random variable.

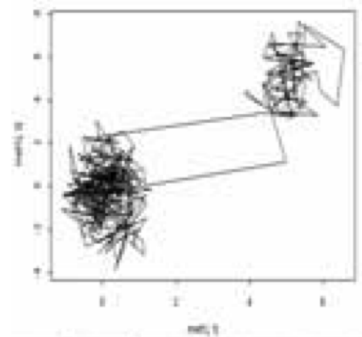
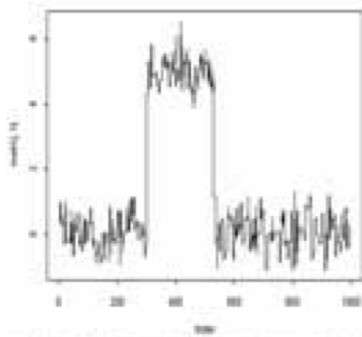
# Parallel Tempering

- This is a fancy M-H algorithm.
- As the name suggests, it consists of running multiple M-H chains in parallel.

**Motivation:** Random-walk Metropolis algorithm makes moves that are very “local” and can therefore fail if our posterior pdf has isolated islands of probability. For example, consider

$$p(\theta_1, \theta_2) \propto e^{-\frac{1}{2} \left( \frac{x^2}{0.25} + \frac{y^2}{2} \right)} + 2 e^{-\frac{1}{2} \left( \frac{(x-5)^2}{0.25} + \frac{(y-5)^2}{2} \right)}.$$

Here is the behavior of the random-walk Metropolis algorithm in this case:



- Suppose we wish to obtain samples from the target density  $p(\boldsymbol{\theta})$ ,  $\boldsymbol{\theta} \in \mathbb{R}^d$ .

- Let us adopt a statistical-mechanics representation of  $p(\boldsymbol{\theta}_1)$ :

$$p(\boldsymbol{\theta}) \propto \exp[-h(\boldsymbol{\theta})/1]$$

and define a family of pdfs:

$$p_i(\boldsymbol{\theta}_i) \propto \exp \left[ - \underbrace{h(\boldsymbol{\theta}_i)}_{\text{energy function}} / t_i \right], \quad t_i > 0.$$

- In general, we may be interested in sampling from  $p_i(\boldsymbol{\theta}_i)$  assuming that  $p_i(\boldsymbol{\theta}_i)$  is a valid pdf, i.e. that  $\int \exp[-h(\boldsymbol{\theta}_i)/t_i] d\boldsymbol{\theta}$  is finite.
- Note that  $h(\mathbf{u}) \leq h(\mathbf{v}) \iff p_i(\mathbf{u}) \geq p_i(\mathbf{v})$ . Therefore, low-energy values correspond to “good” or “high-probability” samples.
- Consider a *temperature ladder* (just a decreasing sequence of positive numbers):

$$t_1 > t_2 > \dots > t_N > 0$$

where  $t_N = 1$ .

- Let us extend the sample space:

$$\boldsymbol{\vartheta} \triangleq [\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_i^T, \dots, \boldsymbol{\theta}_N^T]^T \in \mathbf{R}^{Nd}.$$



- Terminology:

- *population* or *state of the chain*:

$$(\boldsymbol{\theta}_1, t_1; \dots; \boldsymbol{\theta}_i, t_i; \dots \boldsymbol{\theta}_N, t_N).$$

- *i*th *chromosome*:  $\boldsymbol{\theta}_i$ .

- Modified target density:

$$p(\boldsymbol{\vartheta}) \propto \prod_{i=1}^N p_i(\boldsymbol{\theta}_i).$$

Note that

$$p_N(\cdot) = p(\cdot).$$

# Parallel Tempering: Algorithm

**(P)**arallel **(T)**empering consists of two types of moves:

- **M-H update (local move, mutation move)**

- Apply M-H updates to individual chains at different temperature levels (*i.e. to the chromosomes*).

- **Exchange update (global move, random-exchange move)**

- Propose to swap the states of the chains at two neighboring temperature levels (*i.e. two neighboring chromosomes*).

# Mutation Moves

- Choose  $i \in \{1, 2, \dots, N\}$  using some pmf  $q_{\text{mutation}}(i | \boldsymbol{\vartheta})$ .
- In the case of a simple random-walk Metropolis sampler, use

$$(\boldsymbol{\theta}_i)_\star = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i$$

where  $\boldsymbol{\epsilon}_i$  is suitably chosen from a symmetric zero-mean proposal distribution  $J_i(\cdot | \boldsymbol{\theta}_i)$ ,  $i = 1, 2, \dots, N$ ; for example, a popular choice that we mentioned earlier is

$$J_i(\cdot | \boldsymbol{\theta}_i) = \mathcal{N}\left(\boldsymbol{\theta}_i, \underbrace{V_i}_{\text{tuning parameter}}\right).$$

(Here, One can also apply block- or coordinate-wise Gibbs, a slice sampler, or use a general M-H step on  $\boldsymbol{\theta}_i$ .)

Define

$$\boldsymbol{\vartheta}_\star = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, (\boldsymbol{\theta}_i)_\star, \dots, \boldsymbol{\theta}_N).$$

Accept  $\boldsymbol{\vartheta}_\star$  with probability  $\min\{1, r\}$ , where

$$r = \frac{p_i((\boldsymbol{\theta}_i)_\star) q_{\text{mutation}}(i | \boldsymbol{\vartheta}_\star)}{p_i(\boldsymbol{\theta}_i) q_{\text{mutation}}(i | \boldsymbol{\vartheta})}.$$

## Proof.

$$\begin{aligned} r &= \frac{p(\boldsymbol{\vartheta}_\star) J(\boldsymbol{\vartheta} | \boldsymbol{\vartheta}_\star)}{p(\boldsymbol{\vartheta}) J(\boldsymbol{\vartheta}_\star | \boldsymbol{\vartheta})} \\ &= \frac{[\prod_{j=1, j \neq i}^N p_j(\boldsymbol{\theta}_j)] \cdot p_i((\boldsymbol{\theta}_i)_\star) \cdot q_{\text{mutation}}(i | \boldsymbol{\vartheta}_\star) \cdot J_i(\boldsymbol{\theta}_i | (\boldsymbol{\theta}_i)_\star)}{[\prod_{j=1, j \neq i}^N p_j(\boldsymbol{\theta}_j)] \cdot p_i(\boldsymbol{\theta}_i) \cdot q_{\text{mutation}}(i | \boldsymbol{\vartheta}) \cdot J_i((\boldsymbol{\theta}_i)_\star | \boldsymbol{\theta}_i)} \\ &= \frac{p_i((\boldsymbol{\theta}_i)_\star) q_{\text{mutation}}(i | \boldsymbol{\vartheta}_\star)}{p_i(\boldsymbol{\theta}_i) q_{\text{mutation}}(i | \boldsymbol{\vartheta})}. \end{aligned}$$

Here, the  $J_i(\cdot | \cdot)$  terms cancel out because we have employed a Metropolis sampler. In general, they may not cancel out.  $\square$

## Comments:

- At higher temperature levels  $t_i$ , mutation moves are easily accepted because the distribution  $p_i(\cdot)$  is “flat” and thus “hotter” chains travel around the sample space a lot.
- At lower temperature levels  $t_i$ , mutation moves are rarely accepted because the distribution  $p_i(\cdot)$  is very spiky and hence “colder” chains tend to get stuck around a mode.

In other words, at lower temperatures, mutation does “local” exploration and, since the lowest temperature is the temperature of interest, mutations only do not help  $\implies$  we need to consider “mixing” between different chains.

- The “sticking” nature of the mutation moves at lower temperature levels is not necessarily bad — it fosters “local” exploration.

# Random-exchange Moves

- Choose  $i \in \{1, 2, \dots, N\}$  using  $q_{\text{re}}(i | \boldsymbol{\vartheta}) = \frac{1}{N}$  and select  $j \neq i$  such that

$$q_{\text{re}}(j = 2 | i = 1, \boldsymbol{\vartheta}) = 1, \quad q_{\text{re}}(j = N - 1 | i = N, \boldsymbol{\vartheta}) = 1$$

and, for  $i = 2, 3, \dots, N - 1$ ,

$$q_{\text{re}}(j = i \pm 1 | i, \boldsymbol{\vartheta}) = \frac{1}{2}.$$

(This lengthy description simply describes choosing two neighboring chains.)

- Propose to *exchange*  $\boldsymbol{\theta}_i$  and  $\boldsymbol{\theta}_j$  where  $i$  and  $j$  are neighboring values. Define

$$\boldsymbol{\vartheta}_* = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \underbrace{\boldsymbol{\theta}_j}_{i\text{th place}}, \underbrace{\boldsymbol{\theta}_i}_{j\text{th place}}, \dots, \dots, \boldsymbol{\theta}_N).$$

Accept  $\boldsymbol{\vartheta}_*$  with probability  $\min\{1, r\}$ , where

$$\begin{aligned}
 r &= \frac{p_i(\boldsymbol{\theta}_j) p_j(\boldsymbol{\theta}_i)}{p_i(\boldsymbol{\theta}_i) p_j(\boldsymbol{\theta}_j)} \\
 &= \frac{\exp[-h(\boldsymbol{\theta}_j)/t_i] \cdot \exp[-h(\boldsymbol{\theta}_i)/t_j]}{\exp[-h(\boldsymbol{\theta}_i)/t_i] \cdot \exp[-h(\boldsymbol{\theta}_j)/t_j]} \\
 &= \exp\{[h(\boldsymbol{\theta}_j) - h(\boldsymbol{\theta}_i)] \cdot (1/t_j - 1/t_i)\}.
 \end{aligned}$$

**Proof.**

$$\begin{aligned}
 r &= \frac{p(\boldsymbol{\vartheta}_*) J(\boldsymbol{\vartheta} | \boldsymbol{\vartheta}_*)}{p(\boldsymbol{\vartheta}) J(\boldsymbol{\vartheta}_* | \boldsymbol{\vartheta})} \\
 &= \frac{p_1(\boldsymbol{\theta}_1) \cdots p_i(\boldsymbol{\theta}_j) p_j(\boldsymbol{\theta}_i) \cdots p_N(\boldsymbol{\theta}_N)}{p_1(\boldsymbol{\theta}_1) \cdots p_i(\boldsymbol{\theta}_i) p_j(\boldsymbol{\theta}_j) \cdots p_N(\boldsymbol{\theta}_N)} \\
 &= \frac{p_i(\boldsymbol{\theta}_j) p_j(\boldsymbol{\theta}_i)}{p_i(\boldsymbol{\theta}_i) p_j(\boldsymbol{\theta}_j)}.
 \end{aligned}$$

□

**Comments:**

- If  $i > j$  and  $h(\boldsymbol{\theta}_j) \leq h(\boldsymbol{\theta}_i)$ , then

$$r > 1$$

because  $1/t_j < 1/t_i$ . Therefore, the exchange-move is always accepted in this case.

- In words, “good” (low-energy) samples are brought down the ladder and “bad” (high-energy) samples are brought up the ladder. This move probabilistically transports “good” samples down and “bad” samples up the ladder.
- This move can cause jumps between two widely separated modes, thus random exchange has a “global” nature.



# Parallel Tempering Algorithm

Start from values  $\boldsymbol{\theta}_i^{(0)}$  within the support of  $p(\boldsymbol{\theta})$ , for  $i = 1, 2, \dots, N$ .

Choose a temperature ladder  $\{t_i, i = 1, 2, \dots, N\}$ .

Choose a moves-mixture probability  $q, q \in (0, 1)$ .

Here is one iteration of the parallel-tempering algorithm:

- (i) With probability  $q$ , apply the mutation move  $N$  times on the population;
- (ii) With probability  $1 - q$ , apply the random-exchange move  $N$  times on the population.

Thus, we get draws  $\boldsymbol{\vartheta}^{(0)} \rightarrow \boldsymbol{\vartheta}^{(1)} \rightarrow \boldsymbol{\vartheta}^{(2)} \rightarrow \dots \rightarrow \boldsymbol{\vartheta}^{(t)} \rightarrow \dots$

Upon convergence, we look at  $\boldsymbol{\theta}_N^{(t)}$ ,  $t = t_0 + 1, t_0 + 2, \dots, t_0 + T$  where  $t_0$  defines the burn-in period.