

# Bayesian statistics, simulation and software

## Module 9: More MCMC: Metropolis-Hastings algorithm, burn-in, tuning

Jesper Møller and Ege Rubak

Department of Mathematical Sciences  
Aalborg University

# Recall the propose and accept/reject algorithm

Let  $\pi(x)$  be the *target density*, i.e. the density we want to sample from.

## Accept-reject algorithm

Choose initial value  $x^{(0)}$ .

For  $t = 1, 2, \dots, T$

1. Generate **proposal**:  $y \sim q(x^{(t-1)}, y)$ .
2. Accept proposal with probability:  $a(x^{(t-1)}, y)$   
otherwise reject it.
3. If **accepting**:  $x^{(t)} = y$
4. If **rejecting**:  $x^{(t)} = x^{(t-1)}$

This algorithm generates a realisation of a time homogeneous Markov chain.

# Recall the Metropolis-Hastings algorithm

Provides a specific choice of  $a(x, y)$  when  $q(x, y)$  has been specified:

## Metropolis-Hastings algorithm

- Choose any proposal kernel  $q(x, y)$ .
- Define the *Hastings ratio*

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)},$$

where  $H(x, y) = \infty$  if  $\pi(x)q(x, y) = 0$ .

- The acceptance probability is

$$a(x, y) = \min \{1, H(x, y)\}.$$

Then  $\pi$  is invariant; we need to check irreducibility; even better if also aperiodic — see previous lecture.

## Recall the Metropolis algorithm

This is the special case of the MH-algorithm when the proposal kernel is symmetric:

$$q(x, y) = q(y, x).$$

In this case the Hastings ratio simplifies to

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \frac{\pi(y)}{\pi(x)}.$$

**Example:** The most common example is the random walk Metropolis algorithm, i.e. when  $q(x, y) = q_0(x - y)$  where  $q_0$  is a symmetric function. For example, if the proposal is normally distributed with mean  $x$  and precision  $\tau_p$  (user-specified):

$$q(x, y) = \sqrt{\frac{\tau_p}{2\pi}} \exp\left(-\frac{1}{2}\tau_p(y - x)^2\right).$$

Then,  $q(x, y) = q_0(y - x)$  where  $q_0(z) = \sqrt{\frac{\tau_p}{2\pi}} \exp(-\frac{1}{2}\tau_p z^2)$  is symmetric.

# Burn-in

- Generate  $X^{(0)} \sim \pi_0(x)$ ; its distribution is called the **initial distribution**; it is typically different from  $\pi(x)$ .
- Create an irreducible Markov chain  $X^{(0)}, X^{(1)}, X^{(2)}, \dots$  having  $\pi(x)$  as invariant distribution.
- For small values of  $t$ , the distribution of  $X^{(t)}$  can be quite different from  $\pi(x)$ .
- As a consequence, the sample mean

$$\frac{1}{T} \sum_{t=1}^T X^{(t)}$$

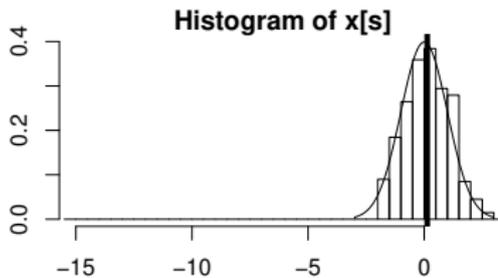
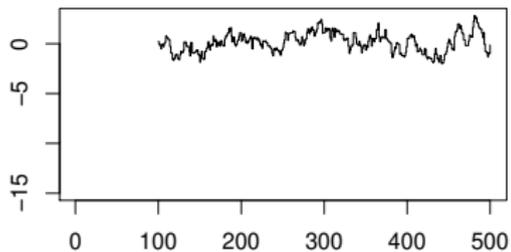
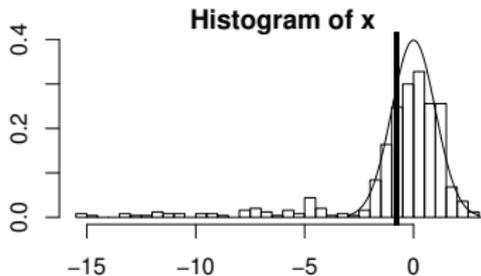
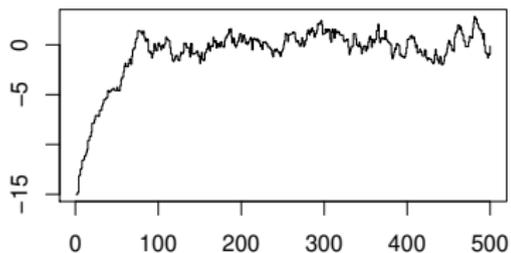
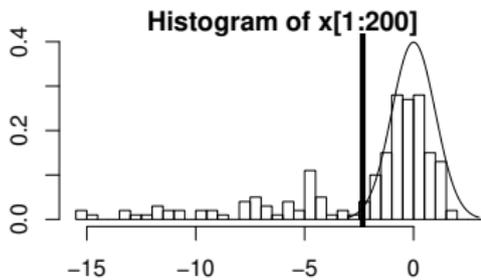
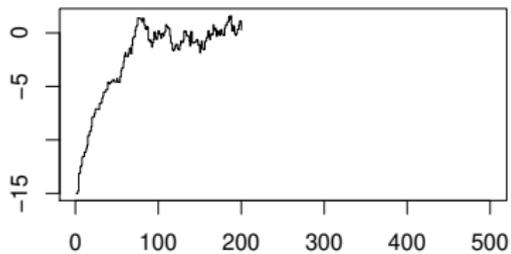
is biased, i.e.  $\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T X^{(t)} \right] \neq \mu$ .

- Instead consider

$$\frac{1}{T} \sum_{t=1}^T X^{(m+t)},$$

where  $m$  is the length of the burn-in.

# The effect of the burn-in



## Variance of the sample mean: IID case

Assume we have independent samples  $X^{(1)}, X^{(2)}, \dots, X^{(T)}$  from  $\pi(x)$ .

Assume  $E[X^{(t)}] = \mu$  and  $Var[X^{(t)}] = \sigma^2$ .

The **sample mean** is

$$\frac{1}{T} \sum_{t=1}^T X^{(t)}$$

and we have the following results:

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T X^{(t)} \right] = \mu$$

$$\text{Var} \left[ \frac{1}{T} \sum_{t=1}^T X^{(t)} \right] = \frac{1}{T} \sigma^2$$

$$T \cdot \text{Var} \left[ \frac{1}{T} \sum_{t=1}^T X^{(t)} \right] = \sigma^2.$$

## Variance of the sample mean: Markov Chain case

Assume  $X^{(1)}, X^{(2)}, X^{(3)}, \dots$  are one-dimensional and form an irreducible Markov chain with invariant density  $\pi(x)$ .

Further, assume that  $X^{(1)} \sim \pi$ . Then  $X^{(t)} \sim \pi$  for  $t = 1, 2, 3, \dots$ , and so  $\mathbb{E}[X^{(t)}] = \mu$  and  $\text{Var}[X^{(t)}] = \sigma^2$  for  $t = 1, 2, 3, \dots$

The expected value of the sample mean is

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T X^{(t)} \right] = \mu.$$

So the expected value of the sample mean is unaffected by the shift from an IID sample to a Markov chain.

## Variance of the sample mean: Markov Chain case

- Regarding the variance, under mild conditions, as  $T \rightarrow \infty$ ,

$$T \cdot \text{Var} \left[ \frac{1}{T} \sum_{t=1}^T X^{(t)} \right] \rightarrow \sigma^2 \left( 1 + 2 \sum_{i=1}^{\infty} \rho_i \right)$$

where

$$\rho_i = \text{Corr}(X^{(t)}, X^{(t+i)}) = \frac{\mathbb{E} [(X^{(t)} - \mu)(X^{(t+i)} - \mu)]}{\sigma^2}$$

is the **lag- $i$  auto-correlation** when assuming  $X^{(t)}$  and  $X^{(t+i)}$  follow  $\pi$ .

- We call  $\sigma^2 (1 + 2 \sum_{i=1}^{\infty} \rho_i)$  the **asymptotic variance**, and  $\tau = 1 + 2 \sum_{i=1}^{\infty} \rho_i$  the **asymptotic correlation**. NB:  $\tau \geq 0$ .
- Trade-off: a small value of  $\tau$  (i.e., negative correlations) seems like a good idea when estimating  $\mu$ , but theory shows that an aperiodic and irreducible MC will then typically convergence slowly to  $\pi$ .
- In fact MH algorithms have often positive correlations (see the next example...)

The following is an example of **tuning** a MH algorithm: Consider a random walk Metropolis algorithm where the proposal kernel is

$$q(x, y) = \sqrt{\frac{\tau_p}{2\pi}} \exp\left(-\frac{1}{2}\tau_p(y-x)^2\right).$$

Here,  $\tau_p$  is a “*user-specified/tuning/algorithm parameter*”, but what is a good choice?

**Example:** On one hand, if the target density is a standard normal,

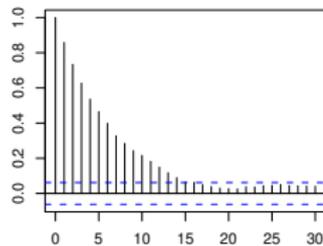
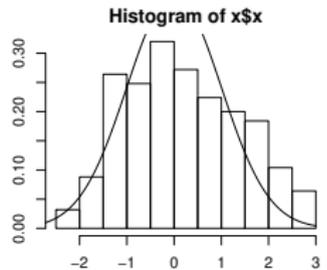
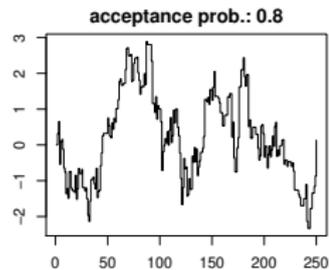
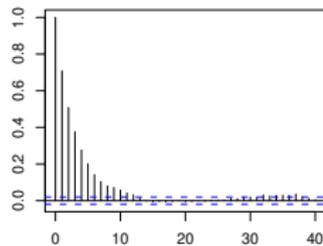
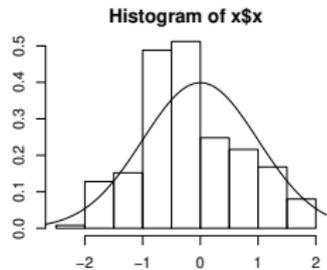
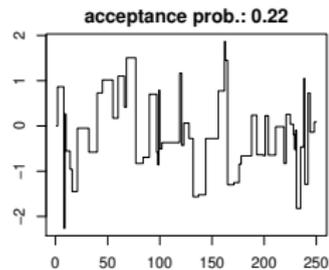
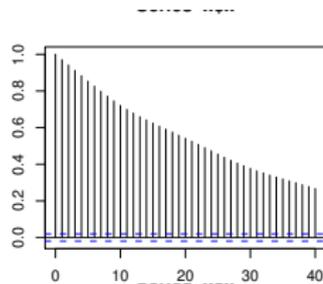
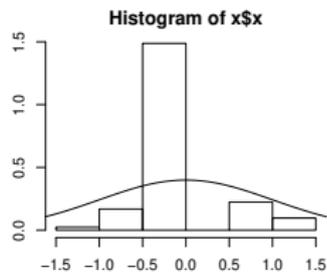
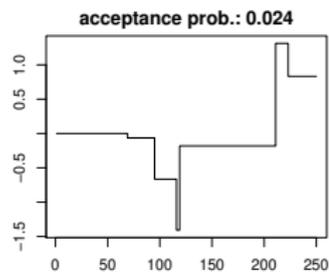
$$\pi(x) = \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

the optimal choice of  $\tau_p$  (in terms of reducing the asymptotic variance) is so that the acceptance probability in average is around 0.4.

On the other hand, if  $\pi(x_1, x_2, \dots, x_k)$  is multivariate normal ( $k$  large), the optimal choice of  $\tau_p$  corresponds to an acceptance probability of 0.234.

Practice for random walk Metropolis: aim at in average 20-40% (or 15-45%) for the acceptance probability.

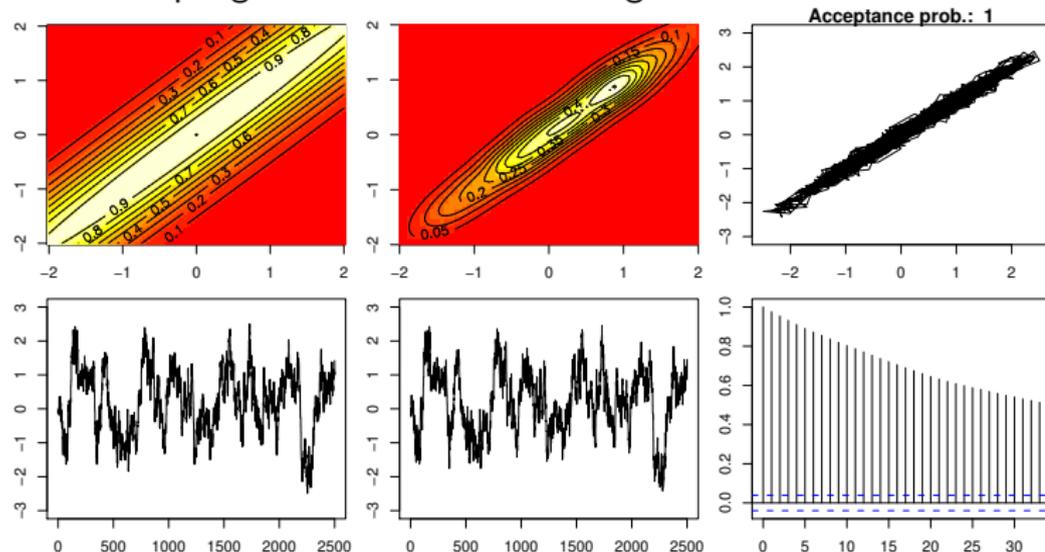
# Tuning: acceptance prob., histogram and auto-correlation



# A bivariate case example

Target:  $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$  (upper left panel).

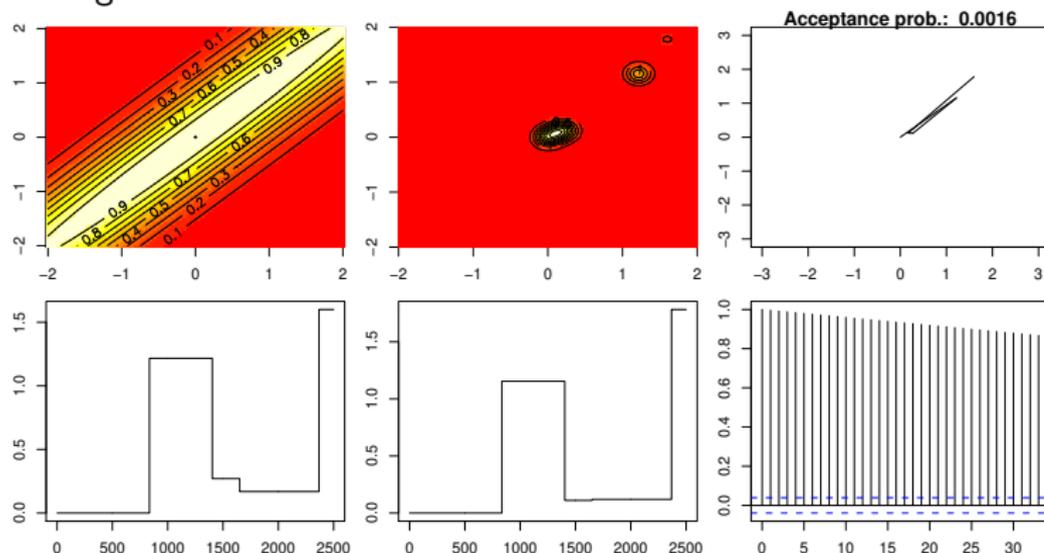
Gibbs sampling: somewhat "slow mixing".



# A bivariate case example

Target:  $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$  (upper left panel).

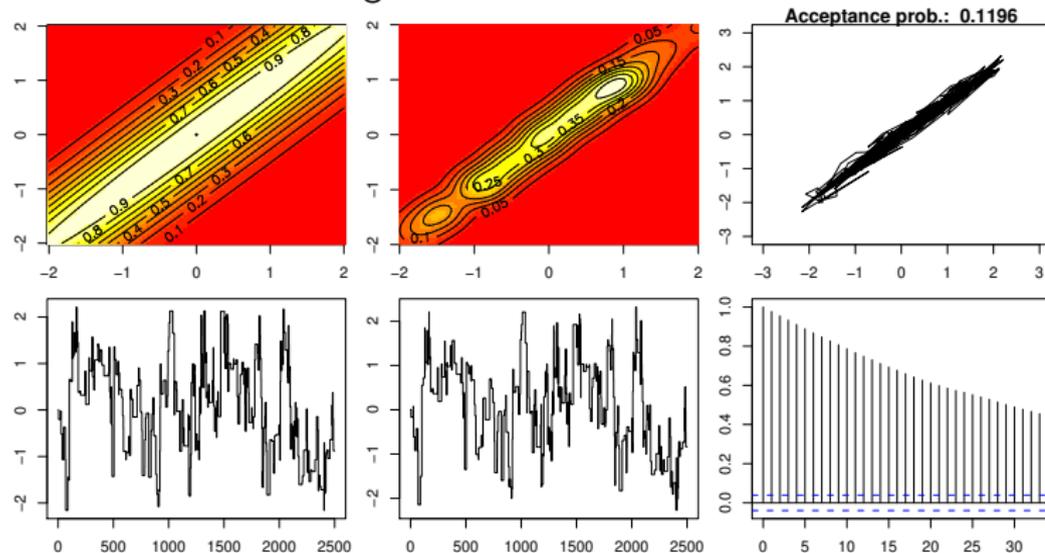
Random walk Metropolis where  $q_0$  is the density of a zero-mean bivariate normal distribution with covariance matrix  $\begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$ : "very poor mixing".



# A bivariate case example

Target:  $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$  (upper left panel).

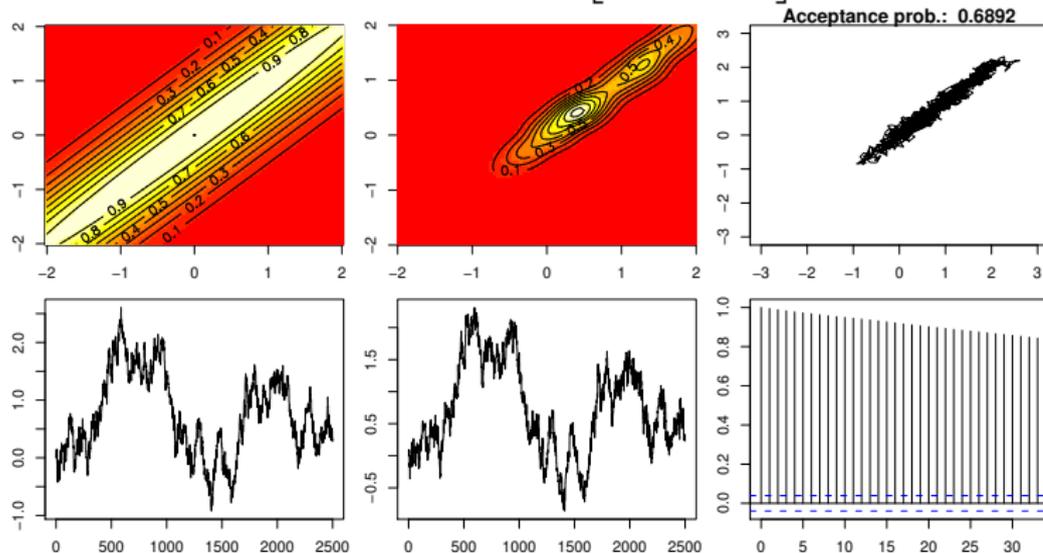
Using instead the covariance matrix  $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ :  
somewhat "slow mixing".



# A bivariate case example

Target:  $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$  (upper left panel).

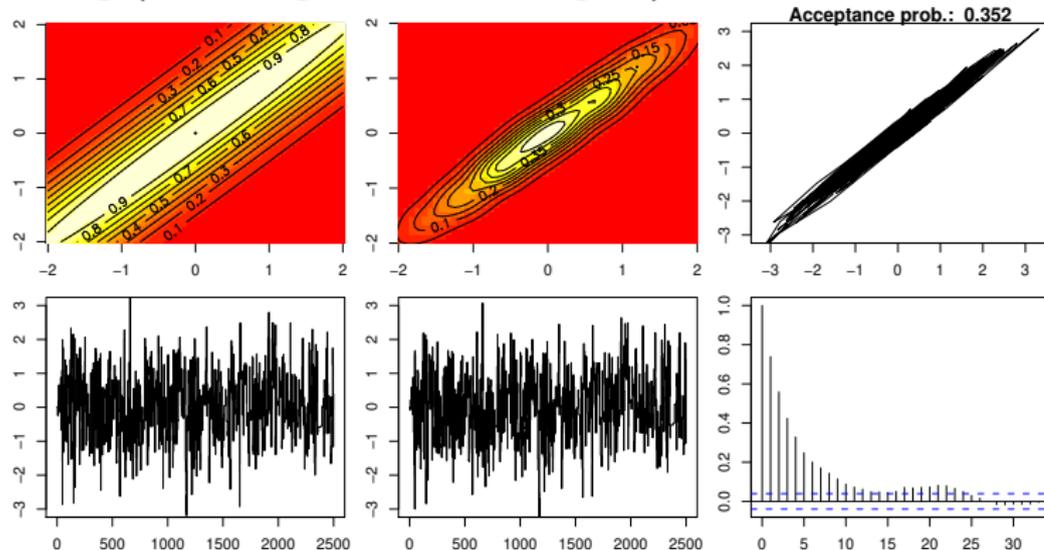
Using instead the covariance matrix  $\begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}$ : "poor mixing".



# A bivariate case example

Target:  $\mathcal{N}\left((0,0), \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}\right)$  (upper left panel).

Using instead the covariance matrix  $\frac{2.38^2}{2} \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$ : "good mixing" (still a longer run would be good).



# Optimum proposal for random walk Metropolis

A purely theoretical result:

Assume target is a  $d$ -dimensional normal:

$$\pi(\mathbf{x}) \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and the proposal is

$$q(\mathbf{x}, \cdot) \sim \mathcal{N}_d(\mathbf{x}, \boldsymbol{\Sigma}_q).$$

Then, as  $d \rightarrow \infty$ , the optimal choice of the proposal variance is

$$\boldsymbol{\Sigma}_q = \frac{2.38^2}{d} \boldsymbol{\Sigma}.$$

## Reminder: The Gibbs sampler

**Aim:** We want to sample  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$  from a density  $\pi(\boldsymbol{\theta})$ , where  $\theta_i \in \Omega_i \subseteq \mathbf{R}^{d_i}$  and  $\pi(\boldsymbol{\theta}) > 0$  for all  $\boldsymbol{\theta} \in \Omega_1 \times \Omega_2 \times \dots \times \Omega_k \subseteq \mathbf{R}^{d_1+d_2+\dots+d_k}$ .

Then we generate an *approximate* sample from  $\pi(\boldsymbol{\theta})$  as follows:

### Gibbs sampler

- Choose initial value  $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)})$ .
- For  $t = 1, 2, \dots, T$ 
  - ▶ For  $i = 1, 2, \dots, k$ 
    1. Generate  $\theta_i^{(t)} \sim \pi(\theta_i | \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \dots, \theta_k^{(t-1)})$

**Question:** What if we cannot generate samples from one or more of the full conditional distributions?

**Solution:** Use a Metropolis-Hastings update instead!

# Metropolis within Gibbs (warning: heavy notation!)

## Metropolis within Gibbs algorithm (MwG)

- Choose initial value  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)})$ .
- For  $t = 1, 2, \dots, T$ 
  - ▶ For  $i = 1, 2, \dots, k$ , cond. on  $\dots = (\theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_i^{(t-1)}, \dots, \theta_k^{(t-1)})$ 
    1. generate proposal  $\theta'_i \sim q(\theta'_i | \dots)$  (NB: may depend on  $\theta_i^{(t-1)}$ )
    2. calculate Hastings ratio

$$H(\dots; \theta'_i) = \frac{\pi(\theta'_i | \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \dots, \theta_k^{(t-1)})}{\pi(\theta_i^{(t-1)} | \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \dots, \theta_k^{(t-1)})} \times \frac{q(\theta_i^{(t-1)} | \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta'_i, \dots, \theta_k^{(t-1)})}{q(\theta'_i | \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_i^{(t-1)}, \dots, \theta_k^{(t-1)})}$$

3. with probability

$$\min \{1, H(\dots; \theta'_i)\}$$

set  $\theta_i^{(t)} = \theta'_i$  (accept) otherwise set  $\theta_i^{(t)} = \theta_i^{(t-1)}$  (reject).

**Remark:** For  $H(\dots; \theta'_i)$  we can work with  $\pi(\cdot | \dots) \propto \dots$

# Metropolis within Gibbs: Comments

- Notice that each component update keeps  $\pi(\boldsymbol{\theta})$  as its invariant distribution, and so the MwG algorithm has  $\pi(\boldsymbol{\theta})$  as its invariant distribution.
- **Special case:** Assume that at some iteration  $i$  of each sweep,

$$\begin{aligned}q(\theta'_i | \dots) &= \pi(\theta'_i | \theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \dots, \theta_k^{(t-1)}) \\ &\propto \pi(\theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta'_i, \theta_{i+1}^{(t-1)}, \dots, \theta_k^{(t-1)})\end{aligned}$$

(i.e., just a Gibbs sampler type update at iteration  $i$ ). Then  $H(\dots; \theta'_i) = 1$ , hence all proposals are accepted at iteration  $i$ .

- So the Gibbs sampler is just the special case of MwG where all proposals are simulations from the full conditionals!
- Irreducibility is *not* automatically fulfilled. In brief, okay if the state space is a product space  $\Omega = \Omega_1 \times \dots \times \Omega_k$  and  $q(\theta'_i | \dots) > 0$  for  $i = 1, \dots, k$ .