

Bayesian Learning

Lecture 8 - Markov Chain Monte Carlo and Metropolis-Hastings

Bertil Wegmann

Department of Computer and Information Science
Linköping University



Lecture overview

- Markov Chain Monte Carlo
- Metropolis-Hastings
- MCMC - efficiency, burn-in and convergence

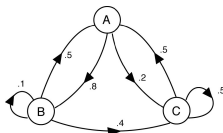
Markov chains

- Let $\mathcal{S} = \{s_1, s_2, \dots, s_k\}$ be a finite set of **states**.
 - ▶ Weather: $\mathcal{S} = \{\text{sunny}, \text{rain}\}$.
 - ▶ School grades: $\mathcal{S} = \{A, B, C, D, E, F\}$
- **Markov chain** is a stochastic process $\{X_t\}_{t=1}^T$ with **state transitions**

$$p_{ij} = \Pr(X_{t+1} = s_j | X_t = s_i)$$

- School grades: $X_1 = C, X_2 = C, X_3 = B, X_4 = A, X_5 = B$.
- **Transition matrix** for weather example

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{pmatrix}$$



Stationary distribution

■ h -step transition probabilities

$$P_{ij}^{(h)} = \Pr(X_{t+h} = s_j | X_t = s_i)$$

■ h -step transition matrix by matrix power

$$P^{(h)} = P^h$$

■ Unique equilibrium distribution $\pi = (\pi_1, \dots, \pi_k)$ if chain is

- ▶ **irreducible** (possible to get to any state from any state)
- ▶ **aperiodic** (does not get stuck in predictable cycles)
- ▶ **positive recurrent** (expected time of returning is finite)

■ Limiting long-run distribution

$$P^t \rightarrow \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \rightarrow \infty$$

Stationary distribution, cont.

■ Limiting long-run distribution

$$P^t \rightarrow \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_k \\ \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_k \end{pmatrix} \text{ as } t \rightarrow \infty$$

■ Stationary distribution

$$\pi = \pi P$$

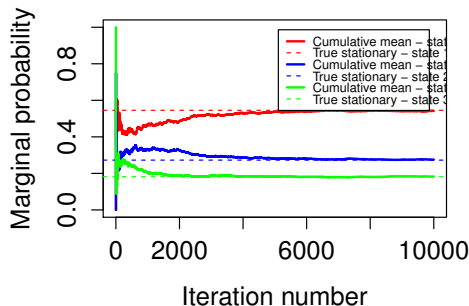
■ Example:

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

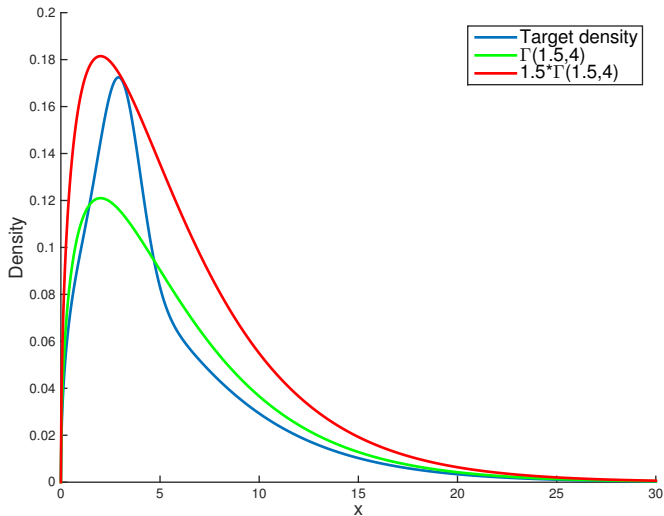
$$\pi = (0.545, 0.272, 0.181)$$

The basic MCMC idea

- Simulate from discrete distribution $p(x)$ when $x \in \{s_1, \dots, s_k\}$.
- **MCMC**: simulate a Markov Chain with a stationary distribution that is exactly $p(x)$.
- How to set up the transition matrix P ? **Metropolis-Hastings**!



Rejection sampling



Random walk Metropolis algorithm

■ **Initialize** $\theta^{(0)}$ and iterate for $i = 1, 2, \dots$

1 **Sample proposal**: $\theta_p | \theta^{(i-1)} \sim N(\theta^{(i-1)}, c \cdot \Sigma)$

2 Compute the **acceptance probability**

$$\alpha = \min \left(1, \frac{p(\theta_p | y)}{p(\theta^{(i-1)} | y)} \right)$$

3 With probability α set $\theta^{(i)} = \theta_p$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Random walk Metropolis, cont.

- Assumption: we can compute $p(\theta_p|y)$ for any θ .
- Proportionality constants in posterior cancel out in

$$\alpha = \min \left(1, \frac{p(\theta_p|y)}{p(\theta^{(i-1)}|y)} \right).$$

- In particular:

$$\frac{p(\theta_p|y)}{p(\theta^{(i-1)}|y)} = \frac{p(y|\theta_p)p(\theta_p)/p(y)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})/p(y)} = \frac{p(y|\theta_p)p(\theta_p)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})}$$

- **Proportional form of posterior is enough!**

$$\alpha = \min \left(1, \frac{p(y|\theta_p)p(\theta_p)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})} \right)$$

Random walk Metropolis, cont.

- Common choices of Σ in proposal $N(\theta^{(i-1)}, c \cdot \Sigma)$:
 - ▶ $\Sigma = I$ (proposes 'off the cigar')
 - ▶ $\Sigma = J_{\hat{\theta}, y}^{-1}$ (propose 'along the cigar')
 - ▶ **Adaptive**. Start with $\Sigma = I$. Update Σ from initial run.
- Set c so average acceptance probability is 25-30%.
- **Good proposal**:
 - ▶ **Easy to sample**
 - ▶ **Easy to compute** α
 - ▶ Proposals should take reasonably **large steps** in θ -space
 - ▶ Proposals should **not be rejected too often**.

The Metropolis-Hastings algorithm

- Generalization when the proposal density is not symmetric.

- Initialize $\theta^{(0)}$ and iterate for $i = 1, 2, \dots$

1 **Sample proposal**: $\theta_p \sim q(\cdot | \theta^{(i-1)})$

2 Compute the **acceptance probability**

$$\alpha = \min \left(1, \frac{p(y|\theta_p)p(\theta_p)}{p(y|\theta^{(i-1)})p(\theta^{(i-1)})} \frac{q(\theta^{(i-1)}|\theta_p)}{q(\theta_p|\theta^{(i-1)})} \right)$$

3 With probability α set $\theta^{(i)} = \theta_p$ and otherwise $\theta^{(i)} = \theta^{(i-1)}$

The independence sampler

- **Independence sampler**: $q(\theta_p | \theta^{(i-1)}) = q(\theta_p)$.

- **Proposal** is **independent of previous draw**.

- **Example**:

$$\theta_p \sim t_\nu(\hat{\theta}, J_{\hat{\theta}, y}^{-1}),$$

where $\hat{\theta}$ and $J_{\hat{\theta}, y}$ are computed by numerical optimization.

- Can be very **efficient**, but has a tendency to **get stuck**.

- Make sure that $q(\theta_p)$ has **heavier tails** than $p(\theta|y)$.

Metropolis-Hastings within Gibbs

- **Gibbs sampling** from $p(\theta_1, \theta_2, \theta_3 | y)$
 - ▶ Sample $p(\theta_1 | \theta_2, \theta_3, y)$
 - ▶ Sample $p(\theta_2 | \theta_1, \theta_3, y)$
 - ▶ Sample $p(\theta_3 | \theta_1, \theta_2, y)$
- When a **full conditional is not easily sampled** we can simulate from it using **MH**.
- Example: at i th iteration, propose θ_2 from $q(\theta_2 | \theta_1, \theta_3, \theta_2^{(i-1)}, y)$. Accept/reject.
- **Gibbs sampling is a special case of MH** when $q(\theta_2 | \theta_1, \theta_3, \theta_2^{(i-1)}, y) = p(\theta_2 | \theta_1, \theta_3, y)$, which gives $\alpha = 1$. Always accept.

The efficiency of MCMC

■ **How efficient** is MCMC compared to iid sampling?

■ If $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$ are **iid** with variance σ^2 , then

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N}.$$

■ Autocorrelated $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)}$ generated by **MCMC**

$$\text{Var}(\bar{\theta}) = \frac{\sigma^2}{N} \left(1 + 2 \sum_{k=1}^{\infty} \rho_k \right)$$

where $\rho_k = \text{Corr}(\theta^{(i)}, \theta^{(i+k)})$ is the autocorrelation at lag k .

■ **Inefficiency factor**

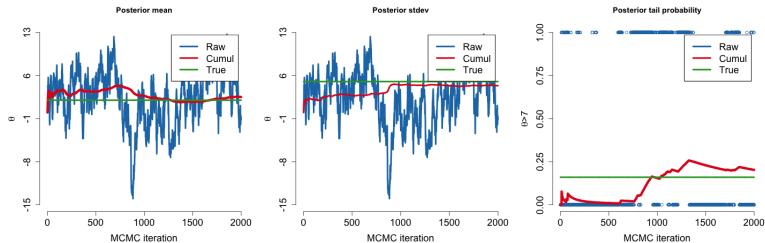
$$\text{IF} = 1 + 2 \sum_{k=1}^{\infty} \rho_k$$

■ **Effective sample size** from MCMC

$$\text{ESS} = N/\text{IF}$$

Burn-in and convergence

- How long **burn-in**?
- How long to **sample** after burn-in?
- **Thinning**? Keeping every h draw reduces autocorrelation.
- **Convergence diagnostics**
 - ▶ Raw plots of simulated sequences (trajectories)
 - ▶ CUSUM plots
 - ▶ ESS
 - ▶ Potential scale reduction factor, \hat{R} .



Burn-in and convergence

