

# ST451 - Lent term

## Bayesian Machine Learning

Kostas Kalogeropoulos

Markov Chain Monte Carlo

# Outline

- 1 Introduction - Motivating Examples
- 2 Markov Chains
- 3 Markov Chain Monte Carlo
- 4 Optional: Metropolis Hasting stationarity proof

# Outline

- 1 Introduction - Motivating Examples
- 2 Markov Chains
- 3 Markov Chain Monte Carlo
- 4 Optional: Metropolis Hasting stationarity proof

# Motivating Examples

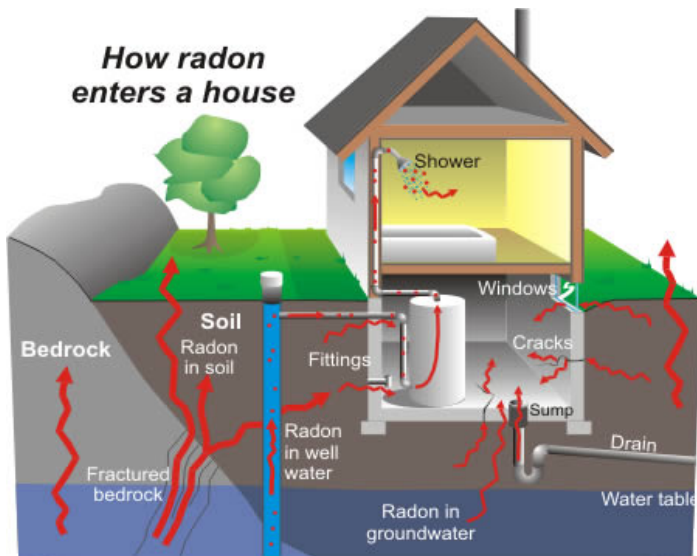
So far we encountered various examples where the posterior was **not available in closed form**, e.g.

- Logistic regression
- Ising model
- Mixtures

We used **approximations** such as Variational Bayes and Laplace.

An alternative option is provided by Monte Carlo where the approximation error can be **controlled** by the user.

## Additional Real World Example: Radon

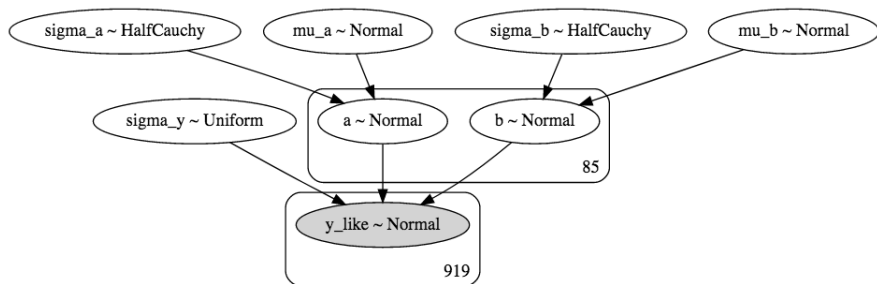


## Hierarchical / Multi-level / Panel data

- Radioactive gas measurements from several households taken from different regions.
- An important predictor is whether the measurement is in the basement or first floor or above.
- The region of the household is also important.
- **Model:**

$$\begin{aligned}y_{ij} &= a_i + b_i X_{ij} + \epsilon_{ij}, \\ \epsilon_{ij} &\stackrel{\text{ind}}{\sim} N(0, \sigma_y^2), \quad a_i \stackrel{\text{ind}}{\sim} N(\mu_a, \sigma_a^2), \quad b_i \stackrel{\text{ind}}{\sim} N(\mu_b, \sigma_b^2), \\ \sigma_y &\sim \text{Uniform}(0, 1), \quad \mu_a \sim N(0, 10^6), \quad \mu_b \sim N(0, 10^6), \\ \sigma_a &\sim \text{HalfCauchy}(5), \quad \sigma_b \sim \text{HalfCauchy}(5)\end{aligned}$$

## DAG of hierarchical model



The model has 175 parameters. Too large for Laplace approximation, Variational Bayes is feasible but still an approximation.

## Bayesian Variable Selection with sparsity

Recall the linear regression model  $y = X\beta + \epsilon$ . When most of the  $X$ 's are **not relevant** with  $y$  it is essential to filter most of them out and come up with **sparse** model.

**Lasso** regression achieves that using the prior

$$\begin{aligned}\beta_i &\sim N(0, \tau^2), \\ \tau &\sim \text{Exp}(\lambda),\end{aligned}$$

but only under the posterior **mode**; the posterior mean is not sparse.

Bayesian sparsity is achieved better by the **spike and slab** priors

$$\begin{aligned}\beta_i &\sim \gamma_i N(0, \tau^2) + (1 - \gamma_i) N(0, \omega^2), \quad \omega \ll \tau \\ \gamma_i &\sim \text{Bernoulli}(\pi)\end{aligned}$$



# Bayesian Variable Selection with sparsity

Nevertheless, the spike and slab choice often leads to **more challenging** computational schemes.

A more convenient approach with very similar behaviour is the **horseshoe** prior. Below is one of its simpler forms

$$\begin{aligned}\beta_i &\sim N(0, \tau^2 \lambda_i), \\ \lambda_i &\sim \text{Cauchy}^+(1), \\ \tau &\sim \text{Cauchy}^+(1),\end{aligned}$$

where  $\text{Cauchy}^+$  denotes the **half-Cauchy** distribution (constrained on the positive real line).

Both horseshoe and spike and slab approaches lead to posteriors that are **not available in closed form**.

# Outline

- 1 Introduction - Motivating Examples
- 2 Markov Chains**
- 3 Markov Chain Monte Carlo
- 4 Optional: Metropolis Hasting stationarity proof

# Markov Chains

We will illustrate the theory for discrete RVs but it also holds for continuous RVs.

Let  $\{x_t, t = 1, \dots, N\}$  be a sequence of (dependent) random variables. They form a **Markov chain** or a **Markov model** if

$$\pi(x_{t+1} | x_1, \dots, x_t) = P(x_{t+1} | x_t)$$

so, we can then write

$$\pi(x_1, \dots, x_N) = \pi(x_1) \prod_{t=1}^{N-1} P(x_{t+1} | x_t)$$



## Stationary / invariant distribution.

The distributions  $P(x_{t+k}|x_t) = T_t(x_t, x_{t+k})$  are called **transition probabilities**. We will focus on cases where they are independent of time and the chain is called **homogeneous**

For a homogeneous Markov chain with transition probabilities  $T(x', x)$ , the distribution  $\pi^*(z)$  is **invariant/stationary** if

$$\pi^*(x) = \sum_{x'} T(x', x) \pi^*(x')$$

The stationary distribution (aka equilibrium) reflects the **long term** behaviour of the Markov Chain.

# Idea 1: Use Markov Chains for simulation

Let  $x$  be a Markov Chain with **transition probability** distribution  $P(x_{t+1}|x_t)$ .

For a given initial value  $x_0$ , we can **simulate**  $x$  in the following way

## Markov Chain Simulation

- 1 **Initialise**. Set  $x_0$ .
- 2 At each time  $t$ , draw the **next** value,  $x_{t+1}$  from  $P(x_{t+1}|x_t)$ .

After a **large**  $t$  all the values of  $X_t$  may be viewed as samples from  $\pi(\cdot)$ . The samples will be **dependent** but still ok for Monte Carlo (unless they are 'too dependent').

## Numerical example of a Markov chain

Consider the Markov chain that is initial value  $x_0$  and **transition** probability

$$x_{t+1}|x_t \sim N(0.5x_t, 1)$$

Let's see its trajectories when started at **two different** starting points.  
(see file 'MarkovChainExample.ipynb')

The **stationary** distribution of  $X$  is the  $N(0, 1.33)$

# Existence of a unique stationary distribution

**Irreducibility:** It is possible to get from any state  $c$  ( $x_t = c$ ) to any state  $d$  at a finite future time  $s$  ( $x_s = d$ ).

**Aperiodicity:** There shouldn't be any loops for all the states.

**Non-null recurrence:** From any state it is possible to return in finite time.

**Ergodicity:** If a chain is non-null recurrent and aperiodic.

**Theorem:** Every irreducible ergodic Markov chain has a limiting distribution, which is equal to  $\pi$  its unique stationary distribution.

# Reversible Markov chains

A chain is **reversible** if it satisfies the **detailed balance** equation:

$$\pi(x_t)P(x_{t+1}|x_t) = \pi(x_{t+1})P(x_t|x_{t+1})$$

Summing over  $x_t$  satisfies the **stationarity** condition

$$\sum_{x_t} \pi(x_t)P(x_{t+1}|x_t) = \sum_{x_t} \pi(x_{t+1})P(x_t|x_{t+1}) = \pi(x_{t+1})$$

For continuous Markov chains replace sums with **integrals**

$$\int \pi(x_t)P(x_{t+1}|x_t)dx_t = \int \pi(x_{t+1})P(x_t|x_{t+1})dx_t = \pi(x_{t+1})$$



# Outline

- 1 Introduction - Motivating Examples
- 2 Markov Chains
- 3 Markov Chain Monte Carlo**
- 4 Optional: Metropolis Hasting stationarity proof

# Markov Chain Monte Carlo: A tale of rediscovery

- First discovered during world war II by **Physicists** in Los Alamos.
- Mainly in Physics but first published in a **Chemistry** journal by Metropolis (1953).
- A publication in **Statistics** by Hastings (1970) was largely unnoticed.
- A special case (Gibbs algorithm) was re-invented for the case of the **Ising** model by Geman and Geman (1984).
- Gelfand and Smith (1990) make the algorithm well-known and Bayesian inference becomes **mainstream**.
- <https://cs.gmu.edu/~henryh/483/top-10.html>

# Main ideas of MCMC

- Construct Markov Chains with the **posterior** as stationary.

**Note:** Possible even if we **only** know the likelihood and the prior.

- Use Markov Chains to **sample** from their stationary distribution.

## Main MCMC algorithms:

- Metropolis Hastings
- Gibbs Sampler
- Hamiltonian MCMC

# Metropolis-Hastings algorithm

From now on switch from  $x$  to  $\theta$  ( $y$  denotes data)

## Metropolis Hastings algorithm

The following algorithm will provide samples from the  $\pi(\theta|y)$

- 1 Initialise  $\theta_0$  at  $t=0$
- 2 Repeat for  $t=0:T-1$ 
  - ▶ Sample a point  $\theta^*$  from  $q(\theta^*|\theta_t)$ .
  - ▶ Set  $\theta_{t+1} = \theta^*$  with probability  $\alpha(\theta_t, \theta^*)$

$$\alpha(\theta_t, \theta^*) = \min \left( 1, \frac{\pi(\theta^*|y)q(\theta_t|\theta^*)}{\pi(\theta_t|y)q(\theta^*|\theta_t)} \right)$$

otherwise set  $\theta_{t+1} = \theta_t$ .

Note that  $\frac{\pi(\theta^*|y)}{\pi(\theta_t|y)} = \frac{f(y|\theta^*)\pi(\theta^*)}{f(y|\theta_t)\pi(\theta_t)}$ . Suffices to know  $\pi(\theta_t|y)$  up to **proportionality**.

# Special cases of Metropolis-Hastings

If we set  $q(\theta^*|\theta_t) = q(\theta^*)$  we get the **Independence sampler**.

$$\alpha(\theta_t, \theta^*) = \min \left( 1, \frac{\pi(\theta^*|y)q(\theta_t)}{\pi(\theta_t|y)q(\theta^*)} \right) = \min \left( 1, \frac{f(y|\theta^*)\pi(\theta^*)q(\theta_t)}{f(y|\theta_t)\pi(\theta_t)q(\theta^*)} \right)$$

Can either perform very well but also poorly.

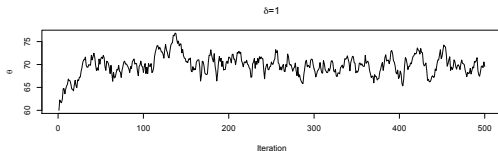
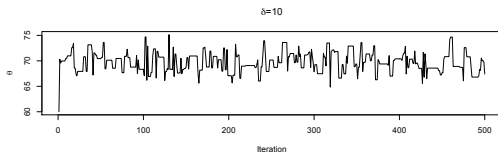
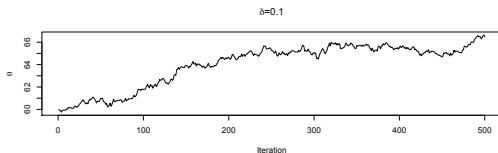
If  $q(\theta^*|\theta_t) = q(\theta_t|\theta^*)$ , e.g.  $N(\theta_t, S_\theta)$ , we get the **Random-walk Metropolis**

$$\alpha(\theta_t, \theta^*) = \min \left( 1, \frac{\pi(\theta^*|y)}{\pi(\theta_t|y)} \right) = \min \left( 1, \frac{f(y|\theta^*)\pi(\theta^*)}{f(y|\theta_t)\pi(\theta_t)} \right)$$

The optimal choice for  $S_\theta$  is a value such that the **acceptance rate** is 0.234.

# Traceplots of Metropolis-Hastings Markov Chains

Convergence and mixing (dependence of the samples) are typically assessed with traceplots. Below we see examples of bad (top), good(middle) and medium (down) cases.



# Gibbs sampler

Suppose that  $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ .

Consider a M-H algorithm where at each iteration we update  $\theta$  as follows: Update **only**  $\theta_1$  first, then **only**  $\theta_2$  and keep going until  $\theta_p$ .

Suppose also that **we know**  $\pi(\theta_i | \theta_{-i}, y)$  for each  $\theta_i$ , where

$$\theta_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p).$$

We can then use  $\pi(\theta_i | \theta_{-i}, y)$ , aka **full conditionals**, as proposals distributions  $q(\theta)$ .

The **acceptance probability will be 1** in all steps (see exercise 1).

# Gibbs Sampler (cont'd)

The Gibbs sampler provides samples from the posterior  $\pi(\theta_1, \dots, \theta_p | y)$

## Gibbs Sampler

- 1 Initialise  $\theta^0 = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)})$
- 2 Repeat for  $t=1:n$ 
  - ▶ Draw  $\theta_1^{(t)}$  from  $\pi(\theta_1 | \theta_2^{(t-1)}, \dots, \theta_p^{(t-1)}, y)$
  - ▶ Draw  $\theta_2^{(t)}$  from  $\pi(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_p^{(t-1)}, y)$
  - ▶ Draw  $\theta_3^{(t)}$  from  $\pi(\theta_3 | \theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_p^{(t-1)}, y)$
  - ...
  - ▶ Draw  $\theta_p^{(t)}$  from  $\pi(\theta_p | \theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{p-1}^{(t)}, y)$



## Example: Normal with $\theta$ and $\sigma^2$ unknown

Given random sample  $y = (y_1, \dots, y_n)$  from  $N(\theta, \sigma^2)$  with  $N(\mu, \tau^2)$  and  $\text{IGamma}(\alpha, \beta)$  as priors. The **posterior** is proportional to

$$\begin{aligned}\pi(\theta, \sigma^2 | y) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta - \mu)^2}{2\tau^2}\right) \\ &\quad (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right)\end{aligned}$$

For  $\pi(\theta | y, \sigma^2)$  gather all the terms involving  $\theta$  and see if you can **identify** the distribution.

$$\begin{aligned}\pi(\theta | y, \sigma^2) &\propto \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta - \mu)^2}{2\tau^2}\right) \\ \dots &\stackrel{D}{=} N\left(\frac{\frac{\sigma^2}{n}\mu + \tau^2\bar{y}}{\tau^2 + \frac{\sigma^2}{n}}, \frac{\tau^2\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}}\right)\end{aligned}$$

## Example: Normal with $\theta$ and $\sigma^2$ unknown (cont'd)

Similarly for  $\pi(|y, \theta)$  we get

$$\begin{aligned}\pi(\sigma^2|y, \theta) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}\right) (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right) \\ &= (\sigma^2)^{-(n/2+\alpha)-1} \exp\left(-\frac{\beta + \frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2}{\sigma^2}\right) \\ &\stackrel{D}{=} \text{IGamma}\left(n/2 + \alpha, \beta + \frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2\right)\end{aligned}$$

A **Gibbs Sampler** initiates  $\theta$  and  $\sigma^2$  and then **alternates** between drawing from the two full conditionals at each iteration.

# Metropolis-Hastings vs Gibbs

Gibbs is generally preferred when it is possible to implement (not available in most cases) as it is **automatic**. It will perform poorly only when  $\theta_i$ 's are highly dependent a-posteriori.

Metropolis-Hastings is **black box** and could perform better than Gibbs in cases of high posterior correlation. But it needs to be tuned; some **adaptive** methods are available.

**Metropolis within Gibbs:** Metropolis-Hastings and Gibbs sampler can be combined by updating each  $\theta_i | \theta_{-i}, y$  with proposals and accept/reject steps.

Metropolis within Gibbs can be used when Gibbs is not available and is hard to tune Metropolis-Hastings.

# Hamiltonian Markov Chain Monte Carlo

Let  $\Phi(\theta) = -\log f(y|\theta) - \log \pi(\theta)$  so that  $\pi(\theta|y) \propto \exp\{-\Phi(\theta)\}$

Extend the **location**  $\theta \in \mathbb{R}^d$  via an auxiliary **velocity**  $v \sim N(0, S)$ ,  $v \perp \theta$ , and consider the **total energy** based on a user-specified covariance  $S$

$$H(\theta) = \Phi(\theta) + \frac{1}{2}v^T S^{-1}v$$

$H(\theta)$  consists of the **potential**  $\Phi(x)$  and the **kinetic energy**  $\frac{1}{2}v^T S^{-1}v$ .

We define the distribution on the  $(\theta, v)$ -space:

$$\pi(\theta, v|y) \propto \exp\{-H(\theta)\} = \exp\{-\Phi(\theta) - \frac{1}{2}v^T S^{-1}v\}$$

# Hamiltonian Dynamics

The **Hamiltonian dynamics** defined on  $\mathbb{R}^{2d}$ , involve gradients and express **preservation of energy**

$$\begin{aligned}\frac{d\theta}{dt} &= v \\ \frac{dv}{dt} &= -S \nabla \Phi(x)\end{aligned}$$

Exact solution of the above equation returns exact samples from  $\pi(\theta, v|y)$ . However **only numerical integrators** are available.

The standard option is the following **leapfrog scheme** ( $L$  and  $h$  need to be specified)

$$\begin{aligned}v_{h/2} &= v_0 - \frac{h}{2} S \nabla \Phi(\theta_0) , \\ \theta_h &= \theta_0 + h v_{h/2} \\ v_h &= v_{h/2} - \frac{h}{2} S \nabla \Phi(\theta_h) ,\end{aligned}$$

# The Hamiltonian MCMC algorithm

The leapfrog scheme is **symmetric** and **volume preserving** but not **energy preserving**. Hence, a **correction** is required, via the following algorithm, to obtain exact samples from  $\pi(\theta|y)$

## Hamiltonian MCMC

- (i) Start with an initial value  $(\theta^{(0)}, v^{(0)}) \sim \otimes_{i=1}^d N(0, 1) \times \pi(\theta)$
- (ii) Given  $\theta^{(k)}$  sample  $v^{(k)} \sim N(0, S)$  and propose and apply  $L$  leapfrog steps to obtain  $(\theta^{(*)}, v^{(*)})$  from  $(\theta^{(k)}, v^{(k)})$
- (iii) Consider

$$a = \min \left( 1, \exp \left\{ -H(x^{(*)}, v^{(*)}) + H(x^{(k)}, v^{(k)}) \right\} \right)$$

- (iv) Set  $\theta^{(k+1)} = x^*$  with probability  $a$ ; otherwise set  $\theta^{(k+1)} = \theta^{(k)}$ .

# Notes on Hamiltonian MCMC

- Also known as Hybrid Monte Carlo.
- It used information from the gradient and results in more 'targeted' proposals.
- It is black-box, i.e. can be applied to any model.
- The parameters  $h$ ,  $L$  and  $S$  need to be specified. This can be done by looking at the history of the chain, but it is not always an easy task.
- Can be implemented via Python and R packages like Stan.

# Today's lecture - Reading

Bishop: 11.1.4 11.2, 11.3 11.5

Murphy: 24.2.1-3 24.3.1-4 24.4.1 24.5.4



# Outline

- 1 Introduction - Motivating Examples
- 2 Markov Chains
- 3 Markov Chain Monte Carlo
- 4 Optional: Metropolis Hasting stationarity proof

## Metropolis Hastings Markov Chains are stationary

**Proposition:** A Markov chain from a Metropolis-Hastings algorithm is reversible. In other words (suppressing the dependency on  $y$ )

$$\pi(\theta_t)P(\theta_{t+1}|\theta_t) = \pi(\theta_{t+1})P(\theta_t|\theta_{t+1})$$

**Proof:** Note that  $P(\theta_{t+1}|\theta_t) = q(\theta_{t+1}|\theta_t)\alpha(\theta_t, \theta_{t+1})$ .

We will consider the three possible cases for  $\pi(\theta_t)q(\theta_{t+1}|\theta_t)$  and  $\pi(\theta_{t+1})q(\theta_t|\theta_{t+1})$  separately.

**Case 1:** If  $\pi(\theta_t)q(\theta_{t+1}|\theta_t) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})$ , then

$$\alpha(\theta_t, \theta_{t+1}) = \frac{\pi(\theta_{t+1})q(\theta_t|\theta_{t+1})}{\pi(\theta_t)q(\theta_{t+1}|\theta_t)} = 1 = \alpha(\theta_{t+1}, \theta_t),$$

so the detailed balance is satisfied with

$$P(\theta_{t+1}|\theta_t) = q(\theta_{t+1}|\theta_t) \text{ and } P(\theta_t|\theta_{t+1}) = q(\theta_t|\theta_{t+1})$$

# Proof of reversibility of Metropolis-Hastings

**Case 2:** If  $\pi(\theta_t)q(\theta_{t+1}|\theta_t) > \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})$ , then  $\alpha(\theta_{t+1}, \theta_t) = 1$  so  $\pi(\theta_{t+1})P(\theta_t|\theta_{t+1}) = \pi(\theta_{t+1})q(\theta_t|\theta_{t+1})$ . But

$$\alpha(\theta_t, \theta_{t+1}) = \frac{\pi(\theta_{t+1})q(\theta_t|\theta_{t+1})}{\pi(\theta_t)q(\theta_{t+1}|\theta_t)}.$$

Hence

$$\begin{aligned}\pi(\theta_t)P(\theta_{t+1}|\theta_t) &= \pi(\theta_t)q(\theta_{t+1}|\theta_t)\alpha(\theta_t, \theta_{t+1}) \\ &= \pi(\theta_t)q(\theta_{t+1}|\theta_t) \frac{\pi(\theta_{t+1})q(\theta_t|\theta_{t+1})}{\pi(\theta_t)q(\theta_{t+1}|\theta_t)} \\ &= \pi(\theta_{t+1})q(\theta_t|\theta_{t+1}) \\ &= \pi(\theta_{t+1})P(\theta_t|\theta_{t+1})\end{aligned}$$

**Case 3:** Similar to Case 2.