

# Week 11: Bayesian Computations

## MATH-517 Statistical Computation and Visualization

Linda Mhalla

2023-12-01

# Bayes' Rule

Let  $X$  be a random variable and  $\theta$  a parameter, considered also a random variable:

$$f_{X,\theta}(x, \theta) = \underbrace{f_{X|\theta}(x | \theta)}_{\text{likelihood}} \underbrace{f_{\theta}(\theta)}_{\text{prior}} = \underbrace{f_{\theta|X}(\theta | x)}_{\text{posterior}} f_X(x).$$

- likelihood = frequentist model ( $\theta$  fixed)
- likelihood & prior = Bayesian model ( $\theta$  random)

Denoting by  $x_0$  the observed value of  $X$ :

$$f_{\theta|X=x_0}(\theta | x_0) = \frac{f_{X|\theta}(x_0 | \theta) f_{\theta}(\theta)}{f_X(x_0)} = \frac{f_{X|\theta}(x_0 | \theta) f_{\theta}(\theta)}{\int f_{X|\theta}(x_0 | \theta) f_{\theta}(\theta) d\theta},$$

which is the Bayes' rule. Rewritten:

$$f_{\theta|X=x_0}(\theta | x_0) \propto f_{X|\theta}(x_0 | \theta) f_{\theta}(\theta),$$

in words:                      posterior  $\propto$  likelihood  $\times$  prior

$\propto$  ... proportional to

# Information update

$X = x_0$  and/or  $\theta = (\psi, \lambda)$  can even be vectors:

$$f_{\psi|X=x_0}(\psi | x_0) \propto \int f_{X|\theta}(x_0 | \theta) f_{\theta}(\psi, \lambda) d\lambda$$

- our original (prior) information (belief) about  $\theta$  was updated by observing  $X = x_0$  into the (marginal) posterior
- this can be applied recursively (when a new  $Y$  independent of  $X$  arrives):

$$\begin{aligned} f_{\theta|X=x,Y=y}(\theta | x_0, y_0) &\propto f_{Y,X|\theta}(x_0, y_0 | \theta) f_{\theta}(\theta) \\ &= f_{Y|\theta}(y_0 | \theta) \underbrace{f_{X|\theta}(x_0 | \theta) f_{\theta}(\theta)}_{\text{old posterior}}, \end{aligned}$$

All available information about  $\theta$  is summarized by the posterior

# Prior Densities

- the prior distribution quantifies the researcher's uncertainty about parameters before observing data
- choice of the prior density is important: it is based on the best available information  $\rightarrow$  subjective
- sometimes use an **improper** prior, which is not a true density (has infinite integral) but for which the posterior is a true density
- often use a **non-informative** proper prior, which inputs only **weak information** ( $\neq$  ignorance) (e.g. normal density with very large (finite) variance, Jeffreys prior based on Fisher information matrix, or uniform prior though not transformation-invariant)
- **conjugate** priors make computations easy as they yield a posterior density of the same family (e.g. beta prior/binomial data  $\rightarrow$  beta posterior or gamma prior/Poisson data  $\rightarrow$  gamma posterior)

**Note:** Empirical Bayes can be used to select parameters of the prior

# Prior Densities

- as sample size increases, the effect of the prior is washed out

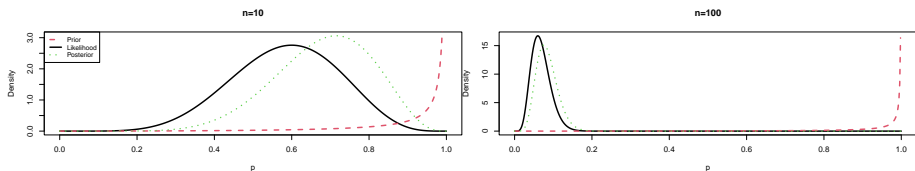
E.g. Bernoulli case

- likelihood:  $\Pr(\mathbf{x}_{1:n}|p) = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}$
- beta prior  $p \sim \text{Beta}(a, b)$ :

$$\Pr(p) \propto p^{a-1} (1-p)^{b-1}$$

on the interval  $(0, 1)$

- posterior:  $\Pr(p|\mathbf{x}_{1:n}) \propto \Pr(\mathbf{x}_{1:n}|p) \Pr(p)$



# Prior Densities: Example

- **Improper Prior**

If  $x \sim \mathcal{N}(\theta, 1)$  and  $f_\theta(\cdot) \equiv \omega$  (constant), then posterior dist. of  $\theta$  is

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

i.e., corresponds to a  $\mathcal{N}(x, 1)$  distribution  $\Rightarrow$  independent of the prior

- **Non-informative (flat) prior**

Consider  $x \sim \mathcal{N}(\theta, 1)$  and  $\theta \sim \mathcal{N}(0, 10)$

$$\begin{aligned} f_{\theta|x}(\theta | x) &\propto f(x | \theta) f_\theta(\theta) \propto \exp \left\{ -\frac{(x - \theta)^2}{2} - \frac{\theta^2}{20} \right\} \\ &\propto \exp \left( -\frac{11\theta^2}{20} + \theta x \right) \propto \exp \left[ -\frac{11}{20} \left\{ \theta - (10x/11) \right\}^2 \right] \end{aligned}$$

and

$$\theta | x \sim \mathcal{N} \left( \frac{10}{11}x, \frac{10}{11} \right)$$

# The Bayesian Approach

- let us denote the data set  $D$ , its realization  $d$ , and  $\theta$  the parameter(s)
- the Bayesian model assumes
  - that nature picks  $\theta$  from the prior distribution  $f_\theta$
  - that nature generates data set  $D = d$  from the likelihood  $f_{D|\theta}$
- the posterior

$$f(\theta \mid D = d) \propto f(d \mid \theta)f(\theta)$$

provides answers for all statistical tasks

- point estimation
- interval estimation
- prediction
- model selection
- hypothesis testing? a matter of choosing priors reflecting the hypotheses
- **Uncertainty:**
  - How much prior belief about  $\theta$  changes in light of data (Bayesian)
  - How estimates vary in repeated sampling from the same population (frequentist)

# Point estimation

**Goal:** a numerical value  $\hat{\theta}$  compatible with the data

Frequentist approach:

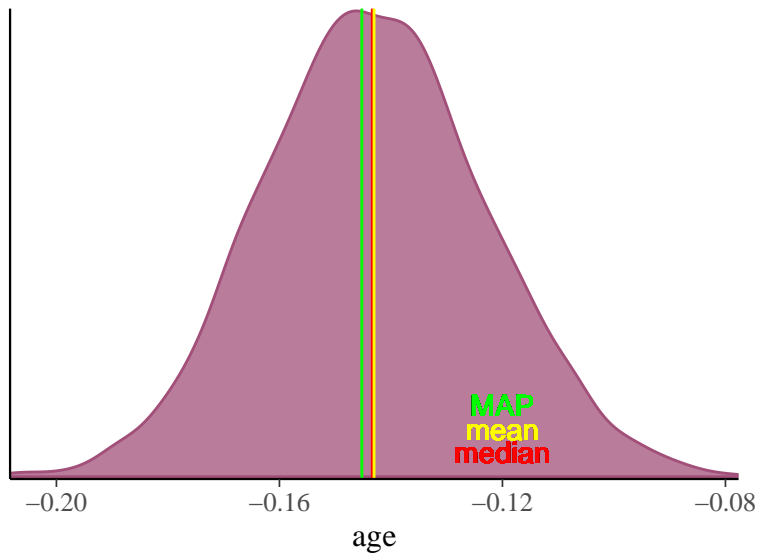
- MLE
- method of moments
- optimization (e.g. penalized least squares), etc.

Bayesian approach:

- MAP - Maximum A Posterior estimate
  - the maximum of the posterior density (close to frequentist MLE)
- posterior mean - the expected value of the posterior
- posterior median
- generally: minimizing the expected loss
  - the expectation is calculated under the posterior
  - e.g., for the squared error loss  $L(\theta, a) = (\theta - a)^2$ , the posterior expected loss  $\int_{\Theta} (\theta - a)^2 dF_{\theta|D}(\theta)$  is minimized at the mean of the posterior distribution;  $|\theta - a|$  yields the posterior median



# Point Estimation



# Interval Estimation

**Goal:** a range of values  $\hat{\theta}$  compatible with the data

Frequentist approach: a confidence interval  $CI_{1-\alpha}$

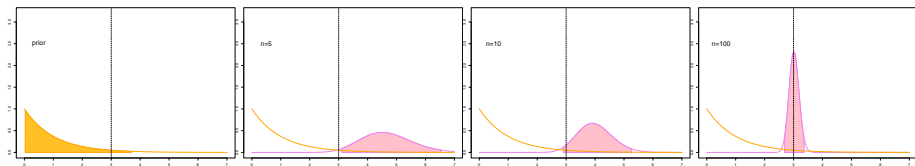
- dual to significance testing
- the probability that the interval contains the true parameter *under replication of the data* is  $1 - \alpha$

Bayesian approach: a credible set  $CR_{1-\alpha}$

- a subset of  $\Theta$  such that  $P(\theta \in CR_{1-\alpha} \mid D) = 1 - \alpha$ 
  - probability calculated under the posterior
- simple interpretation: given the model and data, the probability that the true parameter is in the credible set is  $1 - \alpha$
- many options (just as in the frequentist context), most used: equal-tailed interval and the *highest posterior density set* (narrowest possible)

# Interval Estimation: Equal-tailed Interval

- $CR_{1-\alpha} = [q_{\alpha/2}, q_{1-\alpha/2}]$  where  $q_{\alpha}$  is the  $\alpha$ -quantile of the posterior distribution  $f_{\theta|D}$
- may include values with lower probability than those excluded

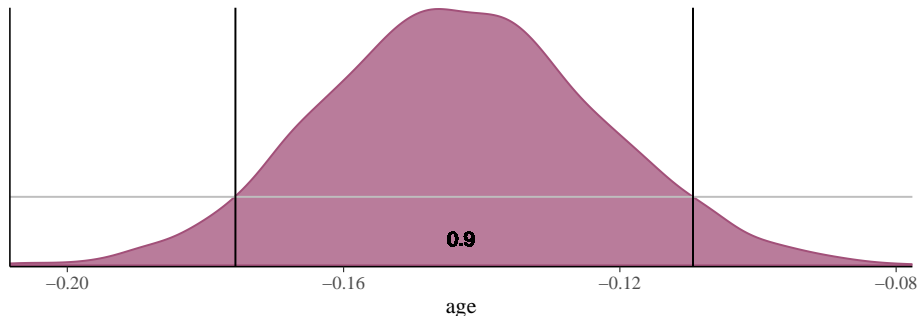


# Interval Estimation: Highest Posterior Density Set

- $\int_{\Theta \cap CR_{1-\alpha}} f_{\theta|D}(\theta | d) d\theta = 1 - \alpha$  such that

$$f_{\theta|D}(\theta | d) \geq f_{\theta|D}(\theta' | d)$$

for all  $\theta \in CR_{1-\alpha}$  and  $\theta' \notin CR_{1-\alpha}$



- not necessarily an interval

# Prediction

**Goal:** posterior prediction, i.e., evaluating or sampling from the posterior predictive distribution  $f_{\tilde{D}|D}$ , where  $D$  is observed data and  $\tilde{D}$  is yet to be observed data

Bayesian approach: prediction = estimation

- assume that likelihood satisfies  $f_{D,\tilde{D}|\theta} = f_{D|\theta} \cdot f_{\tilde{D}|\theta}$ , i.e., new and old data are independent given parameters
- then

$$\begin{aligned}f_{\theta,D,\tilde{D}} &= f_{D|\tilde{D},\theta} \cdot f_{\tilde{D},\theta} = f_{D|\theta} \cdot f_{\tilde{D}|\theta} \cdot f_{\theta} \\ &= f_{\theta,\tilde{D}|D} \cdot f_D\end{aligned}$$

- joint posterior:  $f_{\theta,\tilde{D}|D} = f_{\tilde{D}|\theta} \cdot f_{\theta|D} \Rightarrow$  marginalize out  $\theta$

$$f_{\tilde{D}|D}(\tilde{d} | d) = \int f_{\tilde{D}|\theta}(\tilde{d} | \theta) \cdot f_{\theta|D}(\theta | d) d\theta$$

→ estimated by MC if we can draw from posterior  $f_{\theta|D}$

# Model Selection

**Goal:** decide which element of a set  $\mathcal{M}$  of candidate models fits the data

E.g., the model class  $\mathcal{M}$  can be a mixture of  $K$  Gaussians

Frequentist approach: hypothesis testing, e.g., LRT or constrained inference

Bayesian approach: model selection = estimation (again)

- the data generation process is assumed to have additional level
  - the nature generates a model  $M \in \mathcal{M}$  based on a prior  $f_M$
  - then it generates  $\theta$  conditionally on the model from  $f_{\theta|M}$
  - finally the data are generated conditionally on the model and parameters from  $f_{D|M,\theta}$
- calculate the posterior (now hierarchical):

$$\begin{aligned} f_{D,\theta,M} &= f_{D|\theta,M} \cdot f_{\theta,M} = f_{D|\theta,M} \cdot f_{\theta|M} \cdot f_M \\ &= f_{\theta,M|D} \cdot f_D \end{aligned}$$

posterior:  $f_{\theta,M|D} \propto f_{D|\theta,M} \cdot f_{\theta|M} \cdot f_M$  ... marginalize out  $\theta$  again

## Example: Bayesian Ridge

Consider a Gaussian linear model  $Y = \mathbf{X}\beta + \epsilon$  with  $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{N \times N})$ . Consider the following priors:

- $\beta \sim \mathcal{N}(0, \tau^2 I_{p \times p})$ 
  - $\tau^2$  is a hyperparameter - either fixed or with some hyperprior  $f_{\tau^2}$
- $f_{\sigma^2} \propto 1/\sigma^2$  (improper prior)

Then the posterior for  $\theta = (\beta^\top, \sigma^2, \tau^2)$  is given by

$$f_{\theta|\mathbf{X}, Y}(\beta, \sigma^2, \tau^2 \mid \mathbf{X}, Y) \propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}(Y - \mathbf{X}\beta)^\top (Y - \mathbf{X}\beta)} \frac{1}{\tau^p} e^{-\frac{1}{2\tau^2}\beta^\top \beta} \frac{1}{\sigma^2} f_{\tau^2}(\tau^2)$$

Interestingly, the log-posterior for  $\beta$  is

$$\log f_{\dots}(\beta \mid \mathbf{X}, Y, \sigma^2, \tau^2) \propto -\frac{1}{2\sigma^2}(Y - \mathbf{X}\beta)^\top (Y - \mathbf{X}\beta) - \frac{1}{2\tau^2}\beta^\top \beta$$

so MAP here gives the ridge estimator for  $\lambda = \sigma^2/\tau^2$

# Computational Difficulty

The Bayesian approach above is

- conceptually straightforward and holistic, but
- in practice requires computationally demanding integration
  - the normalization constant
  - marginalization
  - calculating expectations

Possible solutions:

- analytic approximations to the posterior (e.g. Laplace)
- Monte Carlo
  - but the MC techniques we saw already are useful mostly in low-dimensional problems
  - Markov Chain Monte Carlo (MCMC): explore the space in a dependent way, focusing on the important regions



## Section 1

# Markov Chain Monte Carlo (MCMC)

**Goal:** calculate  $\mathbb{E}g(X)$  for some function  $g$  and random variable  $X$

Monte Carlo (MC):

- draw independently  $X_1, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} X$
- approximate  $\mathbb{E}g(X)$  empirically by  $N^{-1} \sum_n g(X_n)$ 
  - works due to LLN

Markov Chain Monte Carlo (MCMC):

- draw  $X^{(1)}, X^{(2)}, \dots, X^{(T)}$  as an *ergodic* Markov Chain with its *stationary distribution* equal to that of  $X$
- approximate  $\mathbb{E}g(X)$  empirically by  $T^{-1} \sum_t g(X^{(t)})$ 
  - works due to the *ergodic theorem* (LLN for Markov sequences)

# Markov Chains

**Definition** (informal): A sequence of random variables  $\{X^{(t)}\}_{t \geq 0}$  with values in  $\mathcal{X} \subset \mathbb{R}^p$  such that

$$X^{(t+1)} \mid X^{(t)}, X^{(t-1)}, \dots, X^{(0)} \sim X^{(t+1)} \mid X^{(t)}$$

is called a discrete-time *Markov chain*

- the conditional distribution  $X^{(t+1)} \mid X^{(t)}$  is given by the *transition kernel*  $k(x, y)$ 
  - for  $X^{(t)} = x$ , the cond. density of  $X^{(t+1)}$  is  $k_x(y) := k(x, y)$
  - $k$  has to meet some conditions on measurability and integrability
  - a Markov chain is fully determined by the transition kernel!
- a distribution  $f$  is called the stationary distribution of a Markov chain associated with a transition kernel  $k$  if

$$\int_{\mathcal{X}} k(x, y) f(x) dx = f(y)$$

If  $f_{t+1}(y) = \int k(x, y) f_t(x) dx = f_t(y)$ , then we stay in the distribution  $f_t$  forever

# Detailed Balance

**Claim:** If the following *detailed balance condition* holds

$$k(x, y)f(x) = k(y, x)f(y)$$

for a distribution  $f$  and a transition kernel  $k$ , then  $f$  is a stationary distribution of the MC associated with  $k$

- $k$  specifies the amount of flow between the points in the domain  $\mathcal{X}$
- detailed balance: the forward flow  $x \rightsquigarrow y$  is equal to the backward flow  $y \rightsquigarrow x$
- equilibrium distribution is preserved: if  $x \sim f$  before a transition, then this is also true afterwards
- let  $f_t$  denote the marginal distribution of  $X^{(t)}$ 
  - $f_0$  is the initial distribution
  - the update  $f_t \rightsquigarrow f_{t+1}$  is governed by  $k$
  - no update  $\Leftrightarrow f_t$  is the stationary distribution  $f$
  - if  $f_0 = f$ , there will never be an update ...  $f_t = f$  for all  $t$

## Definitions:

- A chain is called *irreducible* if any point can be reached (using the kernel) starting from anywhere else

$$\forall u, v \in \mathcal{X}, \quad \exists t \text{ s.t. } P(X^{(t)} = u \mid X^{(0)} = v) > 0$$

- A chain verifying the detailed balance condition (time-reversibility) is called *invariant*
- An irreducible, aperiodic, and invariant chain is called *ergodic*

**Theorem:** If a chain is ergodic then its stationary distribution is the unique invariant distribution and  $f_t \rightarrow f$  for  $t \rightarrow \infty$  regardless of  $f_0$

# Running Monte Carlo via Markov Chains

**Goal:** For an arbitrary starting value  $X^{(0)}$ , construct a chain with a pre-specified stationary distribution  $f$ , typically the posterior  $f_{\theta|D=d}$

- chain = function that generates  $X^{(t+1)}$  depending on  $X^{(t)}$ 
  - the transition kernel  $k$  is in the background
- produce a dependent sample  $X^{(T_0)}, X^{(T_0+1)}, \dots$ , marginally generated from  $f$ , sufficient for most approximation purposes

MCMC is more widely applicable than MC, but what about *mixing*?

- we initialize our chain from  $f_0 \neq f$ 
  - because if we could draw from  $f$ , we would be doing MC instead
  - need to ensure irreducibility
- after a *while*  $f_{T_0} \approx f$  so we have our first draw  $X^{(T_0)} \sim f$ 
  - discard  $X^{(0)}, \dots, X^{(T_0-1)}$  and continue the chain (now stationary)
  - need to ensure invariance (detailed-balance)

**Problem:** How to build a Markov chain with a given stationary distribution? We will see some recipes

# Metropolis–Hastings

**Idea:** construct a candidate new value (a proposal)  $y$  by drawing from arbitrary density  $q(y \mid x)$

- detailed balance requires the right amount of flow between all  $x, y \in \mathcal{X}$
- if there is too much flow  $x \rightsquigarrow y$ , re-map some part of it as  $x \rightsquigarrow x$

Metropolis–Hastings (MH) algorithm:

- **Input:** a proposal density  $q(y \mid x)$ , the target  $f$  (up to a constant)
- **for**  $t = 1, 2, \dots$ , update  $X^{(t-1)}$  to  $X^{(t)}$  by
  - generate  $U^{(t)} \sim q(\cdot \mid X^{(t-1)})$
  - define

$$\alpha(X^{(t-1)}, U^{(t)}) = \min \left\{ 1, \frac{f(U^{(t)})q(X^{(t-1)} \mid U^{(t)})}{f(X^{(t-1)})q(U^{(t)} \mid X^{(t-1)})} \right\}$$

- set  $X^{(t)} := U^{(t)}$  with probability  $\alpha(X^{(t-1)}, U^{(t)})$
- otherwise set  $X^{(t)} := X^{(t-1)}$

(if the proposal is symmetric,  $q$  vanishes from the formula above)

# MH: Convergence Properties

- MH Markov Chain satisfies the detailed balance condition with

$$K(y, x) = \alpha(x, y) q(y | x) + \int \{1 - \alpha(x, \xi)\} q(\xi | x) d\xi \delta_x(y)$$

where  $\delta$  is the Dirac mass

- If  $q(y | x) > 0, \forall x, y$ , then the chain is irreducible
- If

$$P\left(\frac{f(U^{(t)})q(X^{(t-1)} | U^{(t)})}{f(X^{(t-1)})q(U^{(t)} | X^{(t-1)})} \geq 1\right) < 1,$$

that is the event  $\{X^{(t+1)} = X^{(t)}\}$  is possible, then the chain is aperiodic

Thus, under the above two conditions, we have

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T h(X^{(t)}) = \int h(x) df(x)$$



# MH with Random Walk

- use a local perturbation as proposal

$$U^{(t)} = X^{(t-1)} + \epsilon_t,$$

where  $\epsilon_t \sim g$ , independent of  $X^{(t-1)}$

- proposal  $q$  is a symmetric (around 0) density of the form  $q(u - x)$ 
  - e.g.,  $g$  is  $\mathcal{N}(0, \sigma^2)$  hence  $U^{(t)} \sim \mathcal{N}(X^{(t-1)}, \sigma^2)$
  - e.g.,  $g$  is  $\mathcal{U}[-\delta, \delta]$  hence  $U^{(t)} \sim \mathcal{U}[X^{(t-1)} - \delta, X^{(t-1)} + \delta]$
- then,

$$\alpha(X^{(t-1)}, U^{(t)}) = \min \left( 1, \frac{f(U^{(t)})}{f(X^{(t-1)})} \right)$$

# MH with Random Walk

Verifying detailed balance is relatively simple in this case:

- detailed balance:  $k(x, u)f(x) = k(u, x)f(u)$  for  $x \rightsquigarrow u$
- $k(x, u)$  is given implicitly as the mixture of
  - moving away  $x \rightsquigarrow u$  with probability  $\alpha(x, u) = \min \{1, f(u)/f(x)\}$ 
    - $u$  is drawn from  $q(u | x)$  a symmetric density around  $x$
    - equal to
$$k(x, u) = \alpha(x, u)q(u | x) = \alpha(x, u)q(u - x) = \alpha(x, u)q(x - u)$$
  - staying at  $x$  with probability  $1 - \alpha(x, u)$ 
    - i.e.,  $x = u$  ... detailed balance trivially satisfied
- detailed balance:  ~~$q(u | x)$~~   $\alpha(x, u)f(x) = \alpha(u, x)f(u)$   ~~$q(x | u)$~~
- this is trivial since for  $f(x) \neq f(u)$  it is
  - either  $\alpha(u, x) = 1$  and  $\alpha(x, u) = f(u)/f(x)$  leading to

$$\frac{f(u)}{f(x)}f(x) = f(u)$$

- or the other way around

# MH Remarks

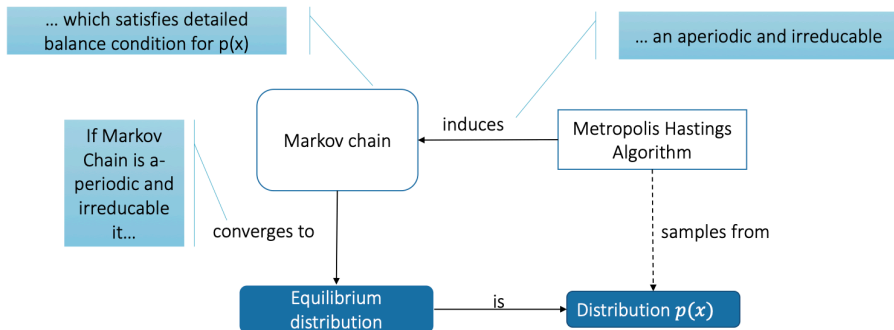
- $f$  is usually a posterior, evaluations needed *up to normalization*
- MH similar in flavor to rejection sampling (RS) in MC
  - but RS needs a majorizing proposal  $g$  to decide accept vs. reject
  - MCMC instead moves vs. stays  $\Rightarrow$  no majorization needed
- never moves to values with  $f(y) = 0$
- the chain  $(X^{(t)})_t$  may take the same value several times in a row, even when  $f$  is a density wrt Lebesgue measure

**Def.:** acceptance rate for MH is the average acceptance probability

$$\bar{\alpha} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \alpha(X^{(t-1)}, U^{(t)})$$

- if  $\bar{\alpha}$  too large, we are probably not exploring the space, mostly staying close with our proposals to where we already were
- if  $\bar{\alpha}$  too small, we have a lot of repeated values in our sample and hence the effective sample size is small even for large  $T$
- 10-50% tends to be a good rate in practice

# MH: Summary



credit: Marcel Lüthi, University of Basel

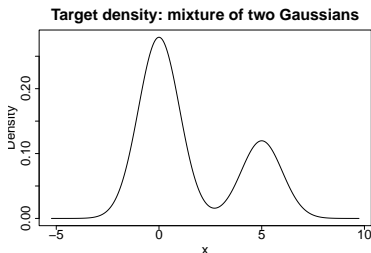
# Example

Consider the MH algorithm with

- a Gaussian mixture target model

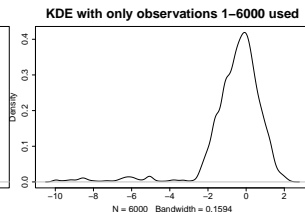
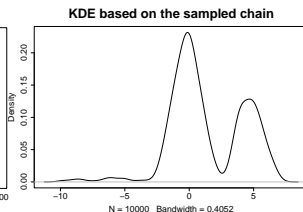
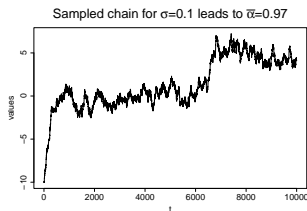
$$f_{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau}(x) = \tau \varphi_{\mu_1, \sigma_1^2}(x) + (1 - \tau) \varphi_{\mu_2, \sigma_2^2}(x)$$

with  $\mu_1 = 1, \mu_2 = 5, \sigma_1 = \sigma_2 = 1$  and  $\tau = 0.7$

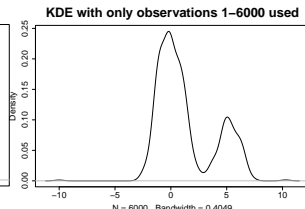
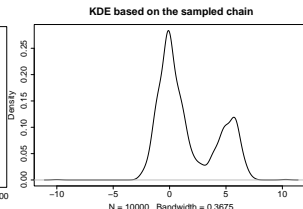
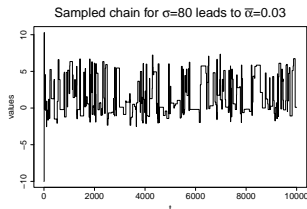


- a Gaussian random walk proposal  $y \sim \mathcal{N}(x, \sigma^2)$ , with  $\sigma = 0.1, 3, 80$
- $x^{(0)} = -10$

# Example

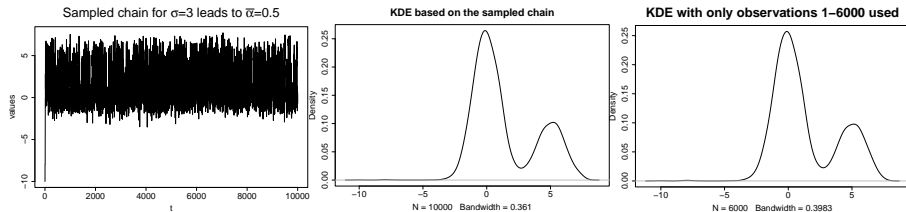


⇒ proposals often accepted but chain moves too slowly



⇒ chain gets stuck for too long

# Example



⇒ seems the best

# References

C. P. Robert & G. Casella (1999) Monte Carlo Statistical Methods

C. P. Robert & G. Casella (2010) Introducing Monte Carlo Methods with  $R$

Gelman & Carlin & Stern & Dunson & Vehtari & Rubin. 2013. Bayesian Data Analysis [updated version](#)



## Assignment 8 [5 %]

Go to [Assignment 8](#) for details