

# Lecture 4 Basics of Bayesian Inference

Shiwei Lan<sup>1</sup>

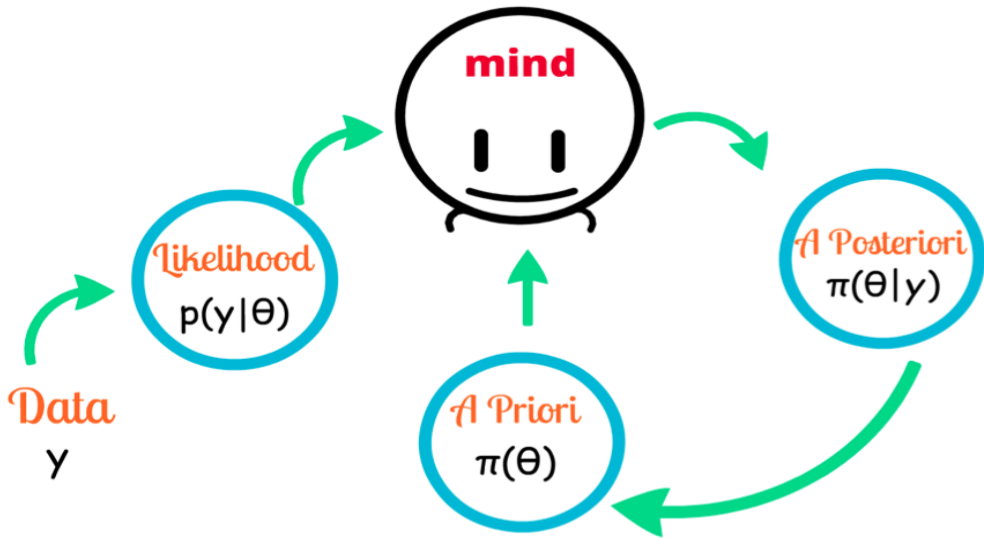
<sup>1</sup>School of Mathematical and Statistical Sciences  
Arizona State University

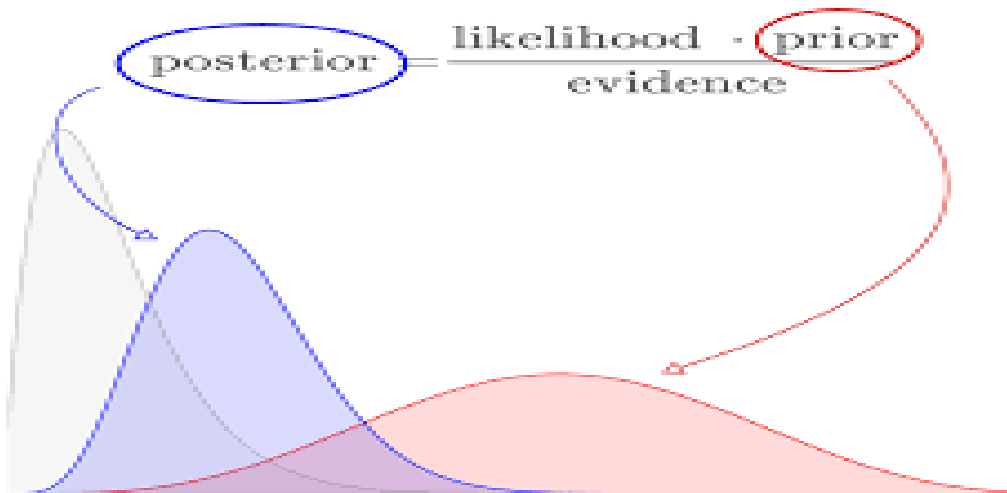
STP598 Spatiotemporal Analysis  
Fall 2020

## 1 Introduction

## 2 Bayesian Inference

## 3 Bayesian Computation





- Probability is subjective: to quantify people's belief of something.
- Bayesian approach provides a coherent framework for combining complex data models and external knowledge or expert opinion.
- It comes with natural uncertainty quantification.
- It enables complex hierarchical models.
- Modern computers make it practical for computation.
- ...

- For two events,  $A$  and  $B$ , *Bayes' theorem* can be simply presented as follows:

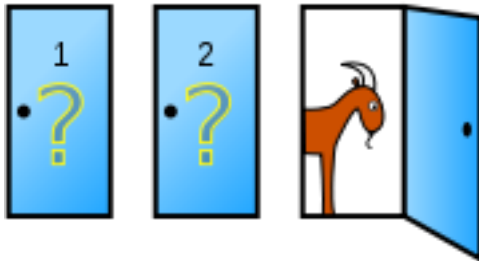
$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

- Also, recall that if  $B = (B_1, B_2, \dots, B_n)$  are a set of events that partition the sample space,  $\Omega$ , using the law of total probability, we have:

$$P(A) = P(A|B_1)P(B_1) + \dots + P(A|B_n)P(B_n)$$
$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i'}^n P(B_{i'})P(A|B_{i'})}$$

- This simple formula which is known as *Bayes' theorem* is the basis of Bayesian analysis. (However, using this theorem does not automatically make you a Bayesian!)

- The problem is based on a TV game show hosted by Monty Hall. In this show, contestants have the chance to win cars. Before each show, a car is put behind one of three closed doors. The other two doors have goats behind them. The contestant is asked to choose one of the three doors. Then, Monty Hall then opens one of the other two doors, which he knows does not have a car behind it. After that, the contestant can choose to switch the the remaining unopened door, or stay with his original selection. The question is: should he switch?



- At the beginning, the car can be behind any of the three doors with equal probability. That is,

$$P(C_1) = P(C_2) = P(C_3) = 1/3$$

- Let's say we choose door number 1, and Monty opens door number 3.
- Now let's write down the conditional probability of opening,  $O_3$ , given the three possibilities (i.e.,  $C_1$ ,  $C_2$ , and  $C_3$ ):

$$P(O_3|C_1) = 1/2$$

$$P(O_3|C_2) = 1$$

$$P(O_3|C_3) = 0$$

- Now using the law of total probability we can find the marginal probability for opening door number 3:

$$P(O_3) = 1/3 \times 1/2 + 1/3 \times 1 + 1/3 \times 0 = 1/2$$



- Using Bayes' theorem, we have:

$$\begin{aligned}P(C_1|O_3) &= \frac{P(C_1)P(O_3|C_1)}{P(O_3)} \\&= \frac{1/3 \times 1/2}{1/2} = 1/3\end{aligned}$$

$$\begin{aligned}P(C_2|O_3) &= \frac{P(C_2)P(O_3|C_2)}{P(O_3)} \\&= \frac{1/3 \times 1}{1/2} = 2/3\end{aligned}$$

- Therefore, probability of winning doubles if we switch.
- You can try this using a penny and three cups.

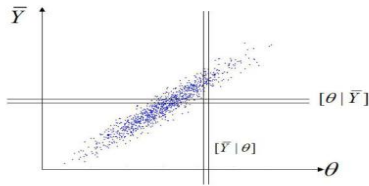
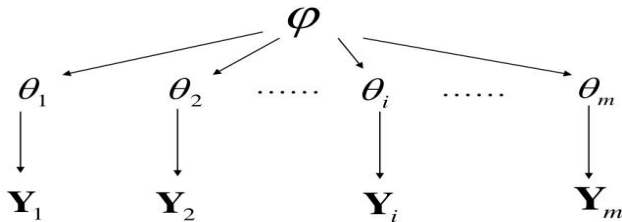
- Given observed data  $y = (y_1, \dots, y_n)$ , we specify the *likelihood* model  $f(y|\theta)$  for a vector of unknown parameter  $\theta = (\theta_1, \dots, \theta_k)$ .
- In Bayesian statistics, we assume  $\theta$  is also a random quantity sampled from a *prior* distribution  $\pi(\theta|\varphi)$ , where  $\varphi$  is a vector of *hyperparameters*.
- If  $\varphi$  is known, inference concerning  $\theta$  is based on its *posterior* distribution

$$p(\theta|y, \varphi) = \frac{p(y, \theta|\varphi)}{p(y|\varphi)} = \frac{f(y|\theta)\pi(\theta|\varphi)}{\int f(y|\theta)\pi(\theta|\varphi)d\theta} \quad (1)$$

- In practice,  $\varphi$  is unknown and given another prior distribution called *hyperprior*, denoted as  $h(\varphi)$ . Then we reduce

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{\int f(y|\theta)\pi(\theta|\varphi)h(\varphi)d\varphi}{\int \int f(y|\theta)\pi(\theta|\varphi)h(\varphi)d\theta d\varphi} \quad (2)$$

## Hierarchical model



- Consider the following linear regression model:

$$y|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n)$$

- $y$  is a column vector of  $n$  outcome observations,  $X$  is an  $n \times (p + 1)$  matrix of predictors with its first column being all 1's.
- $\beta$  is a column vector with  $p + 1$  elements  $(\beta_0, \beta_1, \dots, \beta_p)$  where  $\beta_0$  is the intercept and  $\beta_j$  is the effect of the  $j^{th}$  predictor  $x_j$  on  $y$ .
- In Bayesian analysis, a common prior for parameters are

$$\beta|\mu_0, \Lambda_0 \sim N_{p+1}(\mu_0, \Lambda_0)$$

where  $\mu_0 = (\mu_{00}, \mu_{01}, \dots, \mu_{0p})$  typically set to zero (unless we believe otherwise), and  $\Lambda_0 = \text{diag}(\tau_0^2, \tau_1^2, \dots, \tau_p^2)$  should be sufficiently broad.

- The posterior distributions of  $\beta$  has the following closed form:

$$\begin{aligned}\beta|X, y, \sigma^2 &\sim N(\mu_n, \Lambda_n) \\ \mu_n &= (x_*' \Sigma_*^{-1} x_*)^{-1} x_*' \Sigma_*^{-1} y_* \\ \Lambda_n &= (x_*' \Sigma_*^{-1} x_*)^{-1} \\ x_* &= \begin{pmatrix} X \\ I_{p+1} \end{pmatrix} \quad y_* = \begin{pmatrix} y \\ \mu_0 \end{pmatrix} \quad \Sigma_* = \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \Lambda_0 \end{pmatrix}\end{aligned}$$

- Looking at it this way, the prior plays the role of extra data with  $x_{\beta=I_{p+1}}$ ,  $y_{\beta} = \mu_0$  and the covariance  $\Lambda_0$ .
- That's why Bayesian models do not break down when  $p > n$ .

- Let's take a closer look at the maximum a posterior (MAP)

$$\begin{aligned}\mu_n &= (x'_* \Sigma_*^{-1} x_*)^{-1} x'_* \Sigma_*^{-1} y_* \\ &= (\sigma^{-2} X'X + \Lambda_0^{-1})^{-1} (\sigma^{-2} X'y + \Lambda_0^{-1} \mu_0)\end{aligned}$$

- Let  $\mu_0 = 0$ ,  $\sigma^2 \Lambda_0^{-1} = \lambda I$ , then we have

$$\mu_n = \hat{\beta}^{\text{ridge}} = (X'X + \lambda I)^{-1} X'y$$

- This is exactly the solution to ridge regression!
- If we do not have any informative priors, we can instead assume  $p(\beta|x) \propto 1$ .
- For  $\beta$  this is equivalent (in limit) to taking all  $\tau_j^2 \rightarrow \infty$ .
- The posterior distribution therefore becomes

$$\begin{aligned}\beta|y, \sigma^2 &\sim N(\hat{\beta}, V_\beta \sigma^2) \\ \hat{\beta} &= (X'X)^{-1} X'y, \quad V_\beta = (X'X)^{-1}\end{aligned}$$

- $\hat{\beta}$  is exactly the OLS solution!

- Note, we have

$$\begin{aligned}y|X, \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I_n) \\ \beta|\mu_0, \Lambda_0 &\sim N_{p+1}(\mu_0, \Lambda_0)\end{aligned}$$

- Then a priori, we actually assume

$$y|X, \sigma^2 \sim N(X\mu_0, X\Lambda_0X^T + \sigma^2 I_n) \quad (3)$$

- This can be viewed as a realization of Gaussian process  $y(x) \sim \mathcal{GP}(\mu, \mathcal{C})$ , with the following linear covariance

$$C_{ij} = \text{Cov}(y_i, y_j) = \sum_{k=0}^p x_{ik}x_{jk}\tau_k^2 + \sigma^2\delta_{ij} \quad (4)$$

Bayesian  
Inference

S.Lan

Introduction

Bayesian  
Inference

Bayesian  
Computation

① Introduction

② Bayesian Inference

③ Bayesian Computation



- Point estimation:

- ① posterior mean:  $\hat{\theta} = E[\theta|y]$
- ② posterior median:  $\hat{\theta} : F(\hat{\theta}|y) = \Pr(\theta \leq \hat{\theta}|y) = 0.5$
- ③ posterior mode:  $\hat{\theta} : p(\hat{\theta}|y) = \sup_{\theta} p(\theta|y)$

- Interval estimation:  $(q_L, q_U)$  is a  $100 \times (1 - \alpha)\%$  *credible set* (or *Bayesian confidence interval*) for  $\theta$  if  $\Pr(q_L < \theta < q_U|y) = 1 - \alpha$ .

- ① When the posterior  $\theta|y$  is (nearly) symmetric, then

$$\int_{-\infty}^{q_L} p(\theta|y) d\theta = \int_{q_U}^{\infty} p(\theta|y) d\theta = \alpha/2 \quad (5)$$

- ② *highest posterior density* (HPD) constitutes  $\theta$  having posterior density as large as possible such that

$$1 - \alpha \leq \Pr(C|y) = \int_C p(\theta|y) d\theta \quad (6)$$

- Hypothesis testing is less straightforward. We consider model choice instead.
- Model choice essentially requires specification of the *utility* for a model.
- What are choices of the utility? For what purpose? Explanation? Prediction?
- In the following, we introduce *Deviance Information Criterion (DIC)* and *continuous rank probability score (CRPS)* and other related concepts.

- Given two models  $M_1$  ( $H_0$ ) and  $M_2$  ( $H_A$ ). Associated are two parameter vectors  $\theta_1$  and  $\theta_2$ , with priors  $\pi_i(\theta_i)$  for  $i = 1, 2$  respectively. The marginal distributions of  $Y$  are:

$$p(y|M_i) = \int f(y|\theta_i, M_i)\pi_i(\theta_i)d\theta_i, \quad i = 1, 2 \quad (7)$$

- Then the *Bayes factor* ( $BF$ ), is the ratio of the posterior odds of the  $M_1$  to the prior odds of  $M_1$ , given by Bayes' Theorem as

$$BF = \frac{\Pr(M_1|y)/\Pr(M_2|y)}{\Pr(M_1)/\Pr(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} \quad (8)$$

- This reduces to likelihood ratio for simple-simple model comparison  
 $M_i : \theta = \theta_i$ :  $BF = \frac{f(y|\theta_1)}{f(y|\theta_2)}$

- *Bayesian Information Criterion (BIC)*

$$\Delta\text{BIC} = W - (p_2 - p_1) \log n \quad (9)$$

where  $p_i$  is the number of parameters in model  $M_i$ ,  $i = 1, 2$  and the usual likelihood ratio test statistic  $W = -2 \log \left[ \frac{\sup_{M_1} f(y|\theta)}{\sup_{M_2} f(y|\theta)} \right]$ .

- Schwarz (1978) showed that for nonhierarchical (two-stage) models,  $\text{BIC} \sim -2 \log \text{BF}$  for large  $n$ .
- *Akaike Information Criterion (AIC)*

$$\Delta\text{AIC} = W - 2 \log n \quad (10)$$

- Both AIC and BIC are penalized likelihood ratio model choice criteria; BIC criticizes difference in model dimension more strongly than AIC does.

- Spiegelhalter et al. (2002) propose a generalization of the AIC for hierarchical (3 or more level) models based on the posterior distribution of the *deviance* statistic,

$$D(\theta) = -2 \log f(y|\theta) + 2 \log h(y) \quad (11)$$

where  $f(y|\theta)$  is the likelihood and  $h(y)$  is some standardizing function of data.

- The *complexity* of a model is measured by the effective number of the following parameter  $p_D$

$$p_D = E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}) \quad (12)$$

- *Deviance Information Criterion (DIC)*

$$\text{DIC} = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta}) \quad (13)$$

- Gelfand and Ghosh (1998) propose the *posterior predictive loss* (performance) approach. It focuses on prediction with regard to replicates of the observed data,  $Y_{\ell,rep}, \ell = 1, \dots, n$ .
- The selected models are those that perform well under a so-called *balanced loss* function that penalizes actions both for departure from observed values (“fit”) and for departure from what we expect the replicate to be (“smoothness”).

$$D_k = \frac{k}{k+1}G + P \quad (14)$$
$$G = \sum_{\ell=1}^n (\mu_{\ell} - y_{\ell,obs})^2, \quad P = \sum_{\ell=1}^n \sigma_{\ell}^2$$

where  $\mu_{\ell} = E[Y_{\ell,rep}|y]$  and  $\sigma_{\ell}^2 = \text{Var}[Y_{\ell,rep}|y]$  can be calculated by

$$p(y_{\ell,rep}|y) = \int p(y_{\ell,rep}|\theta_i)p(\theta_i|y)d\theta_i \quad (15)$$

- For prediction task, one can partition the dataset into a fitting (or learning) set and a validation or “hold out” set (20 – 30%). Then apply the criterion to the hold out data after fitting the model to the fitting dataset.
- Gneiting and Raftery (2007) propose the *continuous rank probability score* (CRPS) measure

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(u) - 1(u \geq y))^2 du \quad (16)$$

where  $F$  is the predictive distribution and  $y$  is the observed value.

- Gneiting and Raftery (2007) present a convenient alternative form

$$\text{CRPS}(F, y) = \frac{1}{2} E_F |Y - Y'| = E_F |Y - y| \quad (17)$$

where  $Y$  and  $Y'$  are independent replicates from  $F$ . With samples from  $F$ , we have immediate Monte Carlo integrations.

## 1 Introduction

## 2 Bayesian Inference

## 3 Bayesian Computation



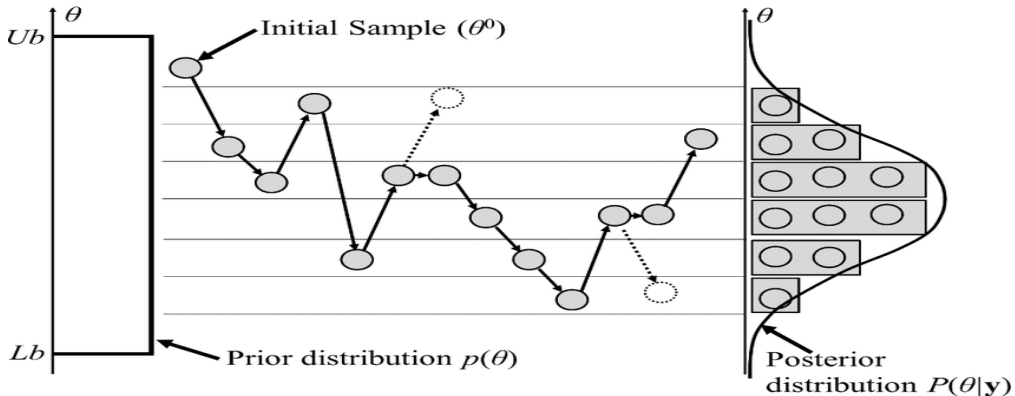
- Suppose that we are interested in sampling from a distribution  $\pi$ . We know the density of  $\pi$  up to a constant, i.e.,  $cf(x)$ .
- We can construct a Markov chain with a transition probability (a.k.a, *proposal distribution*)  $g(x, y)$ .
- Now follow these steps
  - 1 Given our current state  $X^{(n)} = x$ , we propose a new state  $Y^{(n+1)} = y$  according to the transition probability (for example, if we chose  $N(x, 1)$ , we sample from a normal centered at  $x$  with variance 1).
  - 2 Calculated the acceptance probability

$$a(x, y) = \min\left(1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right)$$

- 3 Accept the proposed state  $y$  as the new state with probability  $a(x, y)$  or remain at state  $x$ . That is, sample  $u \sim \text{Unif}(0, 1)$  and set

$$X^{(n+1)} = \begin{cases} y & u < a(x, y) \\ x & \text{otherwise} \end{cases}$$

- The goal for MCMC is to construct a Markov chain with transition kernel  $P(\theta_n, \theta_{n+1})$  such that it converges to the target distribution  $\pi(\cdot)$  in the sense that  $\lim_{n \rightarrow \infty} P^n(\theta_0, A) = \pi(A)$  for any measurable set  $A \subset \Omega$ , regardless of initial point  $\theta_0$ .



- At iteration  $i$ , given our current state  $(x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)})$  we follow these steps:

- 1 Sample  $Y_1^{(i+1)} = y_1$  from the univariate proposal distribution  $g_1(x_1^{(i)}, y_1)$
- 2 Accept this new value and set  $x_1^{(i+1)} = y_1$  with probability

$$a(x_1^{(i)}, y_1) = \min \left[ 1, \frac{p(y_1, x_2^{(i)}, \dots, x_d^{(i)})}{p(x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)})} \right]$$

or reject it and set  $x_1^{(i+1)} = x_1^{(i)}$

- 3 Now sample  $Y_2^{(i+1)} = y_2$  from the univariate proposal distribution  $g_2(x_2^{(i)}, y_2)$ .
- 4 Accept this new value for  $x_2$  with probability

$$a(x_2^{(i)}, y_2) = \min \left[ 1, \frac{p(x_1^{(i+1)}, y_2, \dots, x_d^{(i)})}{p(x_1^{(i+1)}, x_2^{(i)}, \dots, x_d^{(i)})} \right]$$

or reject it and set  $x_1^{(i+1)} = x_1^{(i)}$

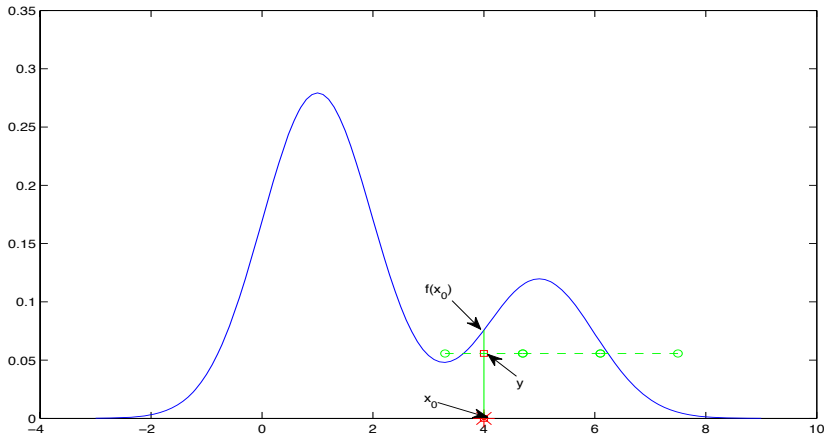
- 5 Continue to update all  $d$  components.

- When we have conditional conjugacy, i.e.  $P(\theta_j|y, \theta_{-j})$  has a closed form, we have the following Gibbs sampler
- At iteration  $i$ , given our current state  $(x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)})$  we cycle through the components one at a time:
  - Sample  $x_1^{(i+1)}$  from the conditional distribution  $P(x_1|x_2^{(i)}, x_3^{(i)}, \dots, x_d^{(i)})$
  - Sample  $x_2^{(i+1)}$  from the conditional distribution  $P(x_2|x_1^{(i+1)}, x_3^{(i)}, \dots, x_d^{(i)})$
  - $\vdots$
  - Sample  $x_j^{(i+1)}$  from the conditional distribution  $P(x_j|x_1^{(i+1)}, \dots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \dots, x_d^{(i)})$
  - $\vdots$
  - Sample  $x_d^{(i+1)}$  from the conditional distribution  $P(x_d|x_1^{(i+1)}, x_2^{(i+1)}, \dots, x_{d-1}^{(i+1)})$

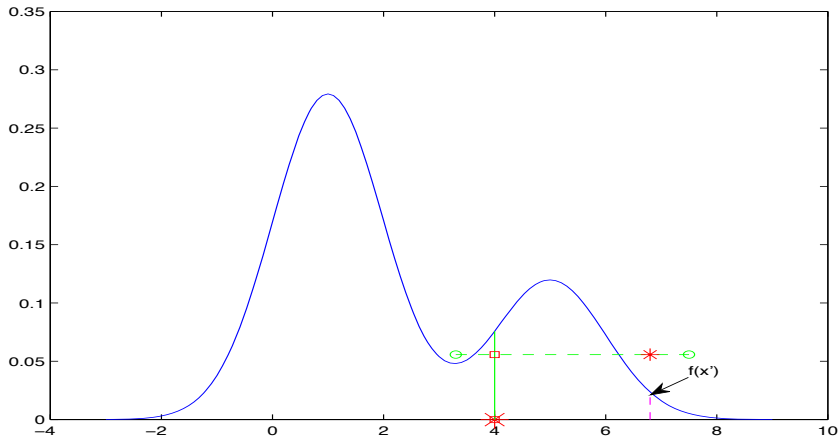
- Note that we are not just proposing anymore, we are directly sampling.
- Or you can look at it as a proposal that will be accepted with probability 1.
- Looking at it this way, the Gibbs sampler is a special case of the Metropolis-Hastings algorithm (note that the transition distribution is not symmetric)

$$\begin{aligned}a(x_j^{(i)}, x_j^{(i+1)}) &= \min \left[ 1, \frac{p(x_j^{(i+1)}, x_{-j}^{(i)})p(x_j^{(i)} | x_{-j}^{(i)})}{p(x_j^{(i)}, x_{-j}^{(i)})p(x_j^{(i+1)} | x_{-j}^{(i)})} \right] \\ &= \min \left[ 1, \frac{p(x_j^{(i+1)} | x_{-j}^{(i)})p(x_{-j}^{(i)})p(x_j^{(i)} | x_{-j}^{(i)})}{p(x_j^{(i)} | x_{-j}^{(i)})p(x_{-j}^{(i)})p(x_j^{(i+1)} | x_{-j}^{(i)})} \right] \\ &= 1\end{aligned}$$

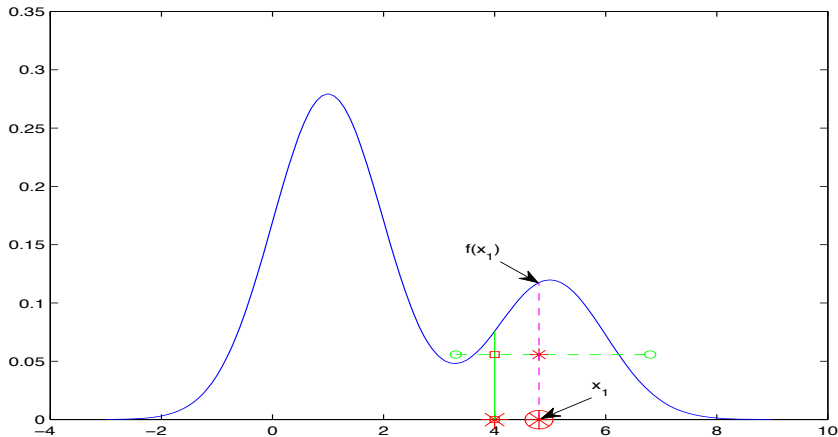
- Sampling  $y \sim \text{Uniform}(0, f(x_0))$  and stepping out (of size  $w$ ) until we reach points outside the area under the density.



- Shrinkage of interval to a point,  $x'$ , which is sampled (uniformly) from the interval but it has  $f(x') < y$ .



- Continuing shrinkage until we reach a point  $x_1$  such that  $y < f(x_1)$ . We accept  $x_1$  as our new sample.





- The following figure (provided in Neal, 2003), shows the stepping out procedure to create an interval  $[L, R]$  around our current point  $x_0$ , with  $y \sim \text{Uniform}(0, f(x_0))$ .

Input:  $f$  = function proportional  
to the density  
 $x_0$  = the current point  
 $y$  = the vertical level defining  
the slice  
 $w$  = estimate of the typical  
size of a slice  
 $m$  = integer limiting the size  
of a slice to  $mw$

Output:  $(L, R)$  = the interval found

$U \sim \text{Uniform}(0, 1)$

$L \leftarrow x_0 - w * U$

$R \leftarrow L + w$

$V \sim \text{Uniform}(0, 1)$

$J \leftarrow \text{Floor}(m * V)$

$K \leftarrow (m - 1) - J$

repeat while  $J > 0$  and  $y < f(L)$ :

$L \leftarrow L - w$

$J \leftarrow J - 1$

repeat while  $K > 0$  and  $y < f(R)$ :

$R \leftarrow R + w$

$K \leftarrow K - 1$

Figure: The stepping out procedure to create interval  $[L, R]$  around  $x_0$ .

- The following figure (provided in Neal, 2003) shows how we can sample a new point from the interval  $[L, R]$  around  $x_0$ .

Input:  $f$  = function proportional  
to the density  
 $x_0$  = the current point  
 $y$  = the vertical level defining  
the slice  
 $(L, R)$  = the interval to sample from  
Output:  $x_1$  = the new point

$\bar{L} \leftarrow L, \bar{R} \leftarrow R$

Repeat:

$U \sim \text{Uniform}(0, 1)$

$x_1 \leftarrow \bar{L} + U * (\bar{R} - \bar{L})$

if  $y < f(x_1)$  and  $\text{Accept}(x_1)$  then  
exit loop

if  $x_1 < x_0$  then  $\bar{L} \leftarrow x_1$   
else  $\bar{R} \leftarrow x_1$

Figure: The shrinkage procedure to create a new sample  $x_1$ .

- Gelman-Rubin-Brooks (1992, 1998) monitor convergence by the estimated *scale reduction factor*

$$\sqrt{\hat{R}} = \sqrt{\left( \frac{N-1}{N} + \frac{m+1}{mN} \frac{B}{W} \right) \frac{df}{df-2}} \quad (18)$$

- Here, we run a small number ( $m$ ) of parallel chains for  $2N$  iterations each.  $B/N$  is the variance between the means from the  $m$  parallel chains.  $W$  is the average of the  $m$  within-chain variances, and  $df$  is the degrees of freedom of an approximating  $t$  density to the posterior distribution.
- The authors show it must approach 1 as  $N \rightarrow \infty$ .
- Geweke's test (1992): test equality of the means for the first segment (usually the first 10% of samples) and the last segment (usually the last 50% of samples) of a Markov chain.
- R packages MCMC and coda.